Wiederholungen in Texten

Segmentieren und Klassifizieren mit vollständigen Substringfrequenzen

D I S S E R T A T I O N

zur Erlangung des akademischen Grades

Dr. phil.
im Fach Germanistische Linguistik

eingereicht an der
Philosophischen Fakultät II
Humboldt-Universität zu Berlin

von
Dipl.-Phys. Felix Golcher

Präsident der Humboldt-Universität zu Berlin:
Prof. Dr. Jan-Hendrik Olbertz

Dekanin der Philosophischen Fakultät II:
Prof. Dr. Helga Schwalm

Gutachterinnen/Gutachter:
1. Prof. Dr. Anke Lüdeling (Humboldt-Universität zu Berlin)
2. Prof. Dr. Marco Baroni (University of Trento)

Tag der mündlichen Prüfung: 28.02.2013

Ich widme diese Arbeit
Islam Bazalaev und Junis Grohmann.

Mein Dank gilt

```
> cat(sample(dankan),sep="\n")
```

Ulf Leser
Verena Harpe
Marc Reznicek
den Autoren der verwendeten Korpora
dem R core team
den Autoren der verwendeten R-Pakete
Anna Renner
Amir Zeldes
Anke Lüdeling
Karsten Tabelow
Yasmin Dalati
Marco Baroni
Marta Lupica Spagnolo
Birte Seffert
Hagen Hirschmann
Juliane Domke

Abstract

This dissertation studies the frequency distribution of all character strings of natural language texts above and below word level regarding to their linguistic and application oriented content.

It has been common practice for a long time now to use sequences of characters or words and frequencies thereof as data basis for corpus and computer linguistic applications. Usually only the most frequent and shortest sequences are taken into account. Since the great majority of character strings of a given text is rare or unique, this restriction excludes a large proportion of existing data from the very start. This negligence of less frequent sequences is motivated on the one hand by the strive for resource efficient, highly performant and simple algorithms and on the other hand by the mostly implicitly made assumption that information contained in less frequent strings is irrelevant.

However, diverse empirical findings raise doubts about the correctness of this position. Over the last 20 years numerous publications (Schenkel et al., 1993; Amit et al., 1994; Ebeling und Pöschel, 1994; Ebeling und Neiman, 1995; Ebeling et al., 1995; Montemurro und Pury, 2002; Golcher, 2005, 2007b; Altmann et al., 2012) have discovered correlations between arbitrarily far distant parts of a text and language independent uniform structures in the repeating text parts, regardless of their length.

This thesis investigates whether these statistical connections and correlations structures, which exist in texts, are exploitable either for new types of technical applications or to deduce facts which are linguistically interpretable beyond the cited basic research. Thus I examine the totality of the character string frequencies in a systematic way, to my knowledge for the first time. To assess their practical linguistic relevance, I review their effectiveness for the solution of different kinds of questions. The chosen research areas should allow qualitative or even quantitative comparisons with published approaches and results. Thus I analyze problems within the following two areas.

First, an algorithm is designed to segment raw unannotated text – a character string – into morphological units (a task also called *Morphological Induction*) based on complete character string frequencies. The aims of this method surpass those of most comparable algorithms since the detected morphological units are further combined at higher levels. In an English text for example not only `accomplish` and `ed` are to be recognized as morphological units. Additionally the algorithm tries to detect that the word `accomplished` forms a unit on a higher level. This thesis succeeds in finding a stable, novel and unsupervised algorithm for this task. It represents a truly rigorous implementation of the old idea that the predictability of the following character drops at borders of language segments (Harris, 1955). Linguistic knowledge beyond that is not used. Although the performance figures of numerous variants of the method have been compared, the same configuration proved to be optimal in all corpora. This feature turns the algorithm into a language independent method for *Morphological Induction*. Through the application of *general* and *generalized linear mixed models* the evaluation reaches a resolution which makes faintest differences between the investigated variants visible. As a whole the technique is capable of giving insight into the morphologies of different languages.

Second, I introduce a *stylometric* method. Stylometry – briefly speaking – is concerned with the quantitative assessment of style. Maybe the best known *stylometric*

task is the automated identification of an author of a text. Besides clarification of authorship many related issues are subsumed under the concept of *stylometry* like the determination of the authors gender or mother tongue. More recent research has shown that modern machine learning techniques on more or less deeply annotated data can reach high and stable performance. I define a text similarity measure on the *complete character string frequencies* of texts and based on it a stylometric classification procedure. The method is evaluated by investigating diverse problems on the basis of different corpora from different languages. The results show that a conceptually simple text similarity measure based on unannotated text can reach or even surpass the high performance of established machine learning techniques, commonly using a much broader data basis.

It is no arbitrary decision to address two seemingly so drastically different questions in one and the same thesis. It is exactly their differences which allow to look at the complete character string frequencies which are the primary research object of this investigation from two different perspectives: *Morphological Induction* mainly concerns the question at which positions a text can be divided into smaller units. This is a question of *local* nature, no matter how wide the considered context is. On the contrary, *stylometry* compares whole texts in order to discover similarities and identify texts by the same author, for example. This task requires a *global* view on the data gathered from a text.

The two main results of this thesis complement each other: First, the complete frequency data of all character sequences of a text are indeed powerful and versatile for possible applications. Second, it turns out that in all used corpora and in view of all investigated tasks and independently of the evaluation method, the logarithmically transformed frequencies are superior to the absolute numbers. That is they yield better morphological segmentations and better stylometric classifications. The logarithm puts different orders of magnitude of frequencies into a balanced relationship. Thus it prevents the very high frequencies of the short character sequences from dominating all computations. On the whole it can be concluded that the longer and less frequent strings contain more relevant and usable information than usually assumed in previous research. This casts a new critical light upon its basic assumptions about the statistical structure of texts.

Zusammenfassung

Diese Dissertation untersucht die Häufigkeitsverteilung sämtlicher in einem natürlichsprachigen Text vorkommenden Zeichenketten sowohl oberhalb als auch unterhalb der Wortebene auf ihren linguistischen und anwendungsbezogenen Informationsgehalt.

Es ist schon lange üblich, Ketten von Zeichen oder Worten und deren Häufigkeiten als Datengrundlage für korpus- und computerlinguistische Anwendungen zu verwenden. Normalerweise werden hierfür nur die häufigsten und kürzesten Ketten betrachtet. Da der überwiegende Teil der Zeichenketten eines Textes selten oder einmalig ist, schließt diese Beschränkung von vornherein den Großteil der insgesamt existierenden Daten aus. Die Nichtbetrachtung der weniger häufigen Zeichenketten motiviert sich einerseits aus dem Streben nach ressourcenschonenden, hoch performanten und möglichst einfachen Algorithmen und andererseits aus der meist nur impliziten Annahme, dass der Informationsgehalt der selteneren Zeichenketten unbedeutend ist.

Vielfältige empirische Erkenntnisse wecken jedoch Zweifel an der Korrektheit dieser Vorstellung. So fördern zahlreiche Veröffentlichungen der vergangenen 20 Jahre (Schenkel et al., 1993; Amit et al., 1994; Ebeling und Pöschel, 1994; Ebeling und Neiman, 1995; Ebeling et al., 1995; Montemurro und Pury, 2002; Golcher, 2005, 2007b; Altmann et al., 2012) Korrelationen zwischen beliebig weit entfernten Teilen eines Textes und sprachunabhängige uniforme Strukturen in den sich wiederholenden Textteilen zu Tage, unabhängig von deren Länge.

In dieser Arbeit wird nun untersucht, ob sich diese in Texten vorhandenen statistischen Zusammenhänge bzw. Korrelationsstrukturen nutzbar machen lassen, sei es für neuartige technische Anwendungen oder um Erkenntnisse abzuleiten, die über die zitierte Grundlagenforschung hinaus linguistisch interpretierbar sind. Daher unterziehe ich die Gesamtmenge der Zeichenkettenhäufigkeiten einer systematischen Untersuchung. Meiner Kenntnis nach ist dies die erste derartige Analyse. Um ihre praktische linguistische Relevanz einer Bewertung zugänglich zu machen, überprüfe ich ihre Wirksamkeit bei der Lösung verschiedener Fragestellungen. Es bieten sich Untersuchungsgebiete an, für die qualitative und womöglich quantitative Vergleichbarkeit mit veröffentlichten Forschungsansätzen und -ergebnissen herstellbar ist. Eine notwendige Voraussetzung ist, dass auf diesen Gebieten bereits vergleichbare Daten verwendet wurden. Daher bearbeite ich Problemstellungen aus den folgenden zwei Themenbereichen.

Zum einen wird ein Algorithmus zur *Morphologischen Induktion* mit Hilfe vollständiger Zeichenkettenhäufigkeiten entwickelt. *Morphologische Induktion* ist die automatisierte Segmentierung von unannotierten Texten – Zeichenketten – in morphologische Einheiten. Das hier vorgestellte Verfahren geht in seiner Zielsetzung über die meisten vergleichbaren Algorithmen hinaus, da die gefundenen morphologischen Einheiten auf höheren Ebenen weiter zusammengefasst werden. So soll in einem englischen Text nicht nur erkannt werden, dass `accomplish` und `ed` morphologische Einheiten darstellen, sondern auch, dass das Wort `accomplished` zusammen auf höherer Ebene wiederum eine Einheit bildet. In dieser Arbeit gelingt es, für diese Fragestellung einen stabilen, neuartigen und unüberwachten Algorithmus vorzustellen. Er stellt eine möglichst konsequente Umsetzung des alten Gedankens dar, dass an den Grenzen *sprachlicher Segmente* die Vorhersagbarkeit der angrenzenden Zeichen abfällt (Harris, 1955). Darüber hinausgehendes sprachliches Wissen

wird nicht implementiert. Obwohl zahlreiche Verfahrensvarianten in ihrer Performanz verglichen werden, erweist sich in allen Korpora ein und dieselbe Konfiguration als optimal. Diese Eigenschaft macht den Algorithmus zu einem sprachunabhängigen Verfahren *Morphologischer Induktion*. Durch die Verwendung *allgemeiner* und *verallgemeinerter linearer gemischter Modelle* erreicht die Evaluation eine Auflösung, die kleinste Unterschiede zwischen den untersuchten Varianten sichtbar macht. Insgesamt ist das Verfahren so geeignet, vergleichende Einblicke in die Morphologien verschiedener Sprachen zu gewinnen.

Zum Anderen stelle ich ein *stilometrisches* Verfahren vor. Stilometrie befasst sich – kurz gesagt – mit der quantitativen Erfassung von Stil. Die wohl bekannteste *stilometrische* Aufgabenstellung ist die automatische Identifizierung des Autors eines Textes. Neben der Klärung der Autorenschaft werden viele verwandte Fragestellungen wie die Bestimmung des Geschlechts oder der Muttersprache eines Verfassers in einem ähnlichen Rahmen untersucht und ebenfalls zur *Stilometrie* gezählt. Die Forschung der neueren Zeit hat ergeben, dass moderne Maschinenlernverfahren auf mehr oder weniger tief annotierten Daten eine hohe und stabile Performanz erreichen können. Ich definiere nun auf den *vollständigen Zeichenkettenhäufigkeiten* von Texten ein Textähnlichkeitsmaß und darauf aufbauend ein stilometrisches Klassifikationsverfahren. Die Methode wird anhand unterschiedlicher Fragestellungen auf einer breiten Datenbasis aus verschiedenen Korpora in verschiedenen Sprachen evaluiert. Es zeigt sich, dass ein konzeptuell einfaches Textähnlichkeitsmaß auf Grundlage unannotierter Texte die hohe Performanz etablierter Maschinenlernverfahren, die darüber hinaus meist auf einer weit breiteren Datenbasis arbeiten, erreichen und unter Umständen übertreffen kann.

Es ist keine willkürliche Entscheidung, zwei anscheinend so unterschiedliche Fragestellungen in ein und derselben Arbeit zu untersuchen. Gerade ihre Verschiedenartigkeit ermöglicht es, die Zeichenkettenhäufigkeiten, denen in dieser Arbeit das Hauptinteresse gilt, aus zwei sehr unterschiedlichen Blickrichtungen zu betrachten: *Morphologische Induktion* behandelt im wesentlichen die Frage, an welcher Stelle ein Text in kleinere Einheiten geteilt werden soll und ist somit notwendigerweise *lokaler* Natur, unabhängig davon, wie weit der betrachtete Kontext ist. Im Gegensatz dazu vergleicht die *Stilometrie* ganze Texte, um Ähnlichkeiten zu erkennen und auf dieser Grundlage zum Beispiel Texte ein und desselben Autors zu identifizieren. Diese Frage erfordert eine *globale* Sicht auf die aus einem Text gewonnenen Daten.

Die zwei Hauptergebnisse der Arbeit ergänzen sich gegenseitig: Zum einen sind die vollständigen Häufigkeitsdaten aller Zeichenketten eines Textes mit Blick auf mögliche Anwendungen tatsächlich mächtig und vielfältig einsetzbar. Andererseits zeigt es sich, dass in allen untersuchten Korpora und in Bezug auf alle untersuchten Fragestellungen und unabhängig von der Evaluationsmethode jeweils die logarithmisch transformierten Häufigkeitsdaten den absoluten Werten überlegen sind, d.h. zu besseren morphologischen Segmentierungen und besseren stilometrischen Klassifikationen führen. Der Logarithmus setzt die verschiedenen Größenordnungen von Häufigkeiten in eine ausgewogene Beziehung und verhindert, dass die sehr großen Häufigkeiten der kurzen Zeichenketten jegliche Berechnung dominieren. Insgesamt kann geschlossen werden, dass in den längeren und selteneren Zeichenketten mehr relevante und nutzbare Information steckt als von der bisherigen Forschung gewöhnlich angenommen wird. Dies lässt deren Grundannahmen über die statistische Struktur von Texten in einem neuen kritischen Licht erscheinen.

Inhaltsverzeichnis

1 Einleitung

1.1 Das Thema der Arbeit und seine Motivation

Diese Dissertation untersucht die Häufigkeitsverteilung sämtlicher in einem natürlichsprachigen Text vorkommenden Zeichenketten sowohl oberhalb als auch unterhalb der Wortebene auf ihren linguistischen und anwendungsbezogenen Informationsgehalt.

Häufigkeiten[1] von Zeichenketten und Wörtern spielten von Beginn an eine Rolle in der korpus- und computerlinguistisch ausgerichteten Forschung. Gilt es beispielsweise in folgendem Satz automatisch zu entscheiden, ob `bear`[2] ein Nomen oder ein Verb ist, so liegt es nahe, in einem Trainingskorpus auszuzählen wie oft das Verb `bear` einem einzelnen `a` folgt im Vergleich zum Nomen `bear` und daraus eine Entscheidung abzuleiten.

> `Isn't that funny that a bear likes honey?` (Milne, 1926)[3]

Ein weiterer Kontext, in dem oft Häufigkeiten von Wörtern oder Zeichenketten verwendet werden, ist die automatische Bestimmung des Autors eines Textes. So könnte das dreimalige Vorkommen der Zeichenkette `t_`[4] mit der Häufigkeit dieses Bigramms in anderen Werken von Milne verglichen werden um Hinweise darauf abzuleiten, ob er tatsächlich deren Autor ist. Detailliertere Beispiele für vergleichbare Forschungsansätze auf Grundlage von Zeichenketten- und Worthäufigkeiten finden sich in den Abschnitten 2.2 und 3.2. In der überwiegenden Zahl der relevanten Arbeiten werden die kurzen, häufigen Ketten und ihre Häufigkeiten in tendenziell großen Textmengen untersucht. In der vorliegenden Arbeit sind Häufigkeiten mit komplementärer Blickrichtung Thema: Zeichenketten unbeschränkter Länge innerhalb relativ kurzer Texte. Im Beispielsatz sind dies 820 Zeichenketten von den Einzelbuchstaben bis zum Satz als Ganzem, von denen die meisten nur ein einziges Mal vorkommen. Meiner Kenntnis nach waren diese vollständigen textstatistischen Daten noch nicht das Thema einer systematischen wissenschaftlichen Untersuchung.

Auf Grundlage der beschriebenen *vollständigen Substringhäufigkeiten* werden in dieser Arbeit vielfältige Fragestellungen aus zwei unterschiedlichen Themenbereichen untersucht.

In Kapitel 2 wird ein Algorithmus zur Segmentierung von unbearbeiteten und unannotierten Texten – Zeichenketten – in morphologische Einheiten (auch bezeichnet als

[1] In Korpus- und Computerlinguistik ist auch *Frequenz* als die direkte Übersetzung des im Englischen gewöhnlich verwendeten Begriffs *frequency* verbreitet und wird entsprechend an zahlreichen Stellen in dieser Arbeit so verwendet. Der Begriff ist allerdings abzugrenzen von der *Frequenz* als dem Kehrwert $f = 1/T$ der Periodendauer T periodischer Bewegungen.

[2] Textbeispiele und Zeichenketten sind in `Schreibmaschinenschrift` gesetzt.

[3] zitiert nach Rowohlt (2005).

[4] Der Unterstrich ersetzt zur besseren Lesbarkeit das Leerzeichen.

Morphologische Induktion) mit Hilfe vollständiger Substringfrequenzen entwickelt. Das Verfahren geht in seiner Zielsetzung über die meisten vergleichbaren Algorithmen hinaus, da die gefundenen morphologischen Einheiten auf höheren Ebenen weiter zusammengefasst werden. So soll in einem englischen Text nicht nur erkannt werden, dass `accomplish` und `ed` morphologische Einheiten darstellen, sondern auch, dass das Wort `accomplished` zusammen auf höherer Ebene wiederum eine Einheit bildet.

Im darauf folgenden Kapitel 3 stelle ich ein *stilometrisches* Verfahren vor. Stilometrie befasst sich – kurz gesagt – mit der quantitativen Erfassung von Stil. Die wohl bekannteste *stilometrische* Aufgabenstellung ist die automatische Identifizierung des Autors eines Textes. Ein prominentes Beispiel ist die Diskussion, ob Christopher Marlowe der Autor (einiger Teile) des gewöhnlich Shakespeare zugeschriebenen Kanons sein könnte (Merriam, 1993; Merriam und Matthews, 1994; Merriam, 2000). Neben der Klärung der Autorenschaft werden viele verwandte Fragestellungen wie die Bestimmung des Geschlechts oder der Muttersprache eines Verfassers in einem ähnlichen Rahmen untersucht und ebenfalls zur *Stilometrie* gezählt.

Es ist keine willkürliche Entscheidung, zwei anscheinend so unterschiedliche Fragestellungen in ein und derselben Arbeit zu untersuchen. Gerade ihre Verschiedenartigkeit ermöglicht es, die vollständigen Substringfrequenzen, denen in dieser Arbeit das Hauptinteresse gilt, aus zwei sehr unterschiedlichen Blickrichtungen zu betrachten: Die Frage, an welcher Stelle ein Text in kleinere Einheiten geteilt werden soll, ist notwendigerweise *lokaler* Natur, unabhängig davon, wie weit der betrachtete Kontext ist. Im Gegensatz dazu vergleicht die Stilometrie ganze Texte, um Ähnlichkeiten zu erkennen und auf dieser Grundlage zum Beispiel Texte ein und desselben Autors zu identifizieren. Diese Frage erfordert eine *globale* Sicht auf die aus einem Text gewonnenen Daten.

Ich verfolge in dieser Arbeit drei Hauptziele: Erstens ermöglicht die Breite der Untersuchungen und ihr in Teilen explorativer Charakter einen ersten Überblick über diese bisher weitgehend unerforschten Daten. Zweitens sollen ihre anwendungsrelevanten Vorteile im Kontext der beiden untersuchten Fragestellungen und im Vergleich zu den jeweils üblicherweise eingesetzten Daten ausgelotet werden. Ein drittes Ziel ist die Gewinnung theoretisch interpretierbarer Schlussfolgerungen, vor allem aus dem Vergleich der Performanz parallel untersuchter Verfahrensvarianten.

Die zwei Hauptergebnisse der Arbeit ergänzen sich gegenseitig: Zum einen sind die vollständigen Häufigkeitsdaten aller Zeichenketten eines Textes mit Blick auf mögliche Anwendungen potentiell mächtig und vielfältig einsetzbar. Andererseits ist die allgemeinste und weitreichendste Beobachtung, die ich aus den vielfältigen empirischen Befunden ableite, dass in den längeren und selteneren Zeichenketten mehr Information steckt als von der bisherigen Forschung gewöhnlich angenommen wird. Dies lässt deren Grundannahmen über die statistische Struktur von Texten in einem neuen kritischen Licht erscheinen.

Die Ausweitung der Datengrundlage von einer Untermenge aller Zeichenketten auf die Gesamtheit der *vollständigen Substringfrequenzen* ermöglicht es, die Berechtigung der oft nur impliziten Argumente für eine Beschränkung auf die gewöhnlich verwendete Untermenge der häufigen und kurzen Zeichenketten zu überprüfen. Diese Überprüfung erscheint lohnend und angebracht, da es ernstzunehmende Hinweise darauf gibt, dass

die vollständigen Substringfrequenzen einzelner Texte sprachwissenschaftlich relevante Information beinhalten, die wesentlich über den Informationsgehalt der kurzen häufigen Ketten hinausgehen. Einige dieser Hinweise seien hier kurz diskutiert.

Von vornherein ist die Beschränkung auf häufige Ereignisse mit Blick auf die allgemeine Struktur sprachlicher Häufigkeitsverteilungen nicht besonders intuitiv (Zipf, 1949; Baayen, 2001; Baroni, 2008): Einige Ereignisse, Wörter zum Beispiel, treten sehr häufig auf, während es eine riesige Masse an extrem seltenen Ereignissen gibt, die zusammen dennoch den Großteil aller auftretenden Ereignisse ausmachen. Daher ist es eine sehr weitgehende Beschränkung auf einen kleinen Teil der insgesamt existierenden Daten, wenn man nur die häufigsten Wörter oder Zeichenketten in eine Analyse einschließt. Auch wenn der Verlust an Information, der mit dieser Beschränkung einhergeht, nicht ohne weiteres exakt quantifizierbar ist und auch sicherlich von der untersuchten Fragestellung abhängt, so gibt es doch zwei Gruppen von Forschungsarbeiten, die konkrete Hinweise darauf erbringen, dass längere Zeichenketten durchaus bisher unbeobachtete und ungenutzte Informationen enthalten.

Starke Indizien ergeben sich aus Forschungen zur allgemeinen funktionalen Abhängigkeit von Korrelationen in Texten über größere Abstände hinweg. Dass es zwischen den Buchstaben eines Textes überhaupt Korrelationen gibt, ist eine triviale Feststellung: In einem deutschen Text ist nach einem `c` die Wahrscheinlichkeit für ein unmittelbar folgendes `h` etwa 26 mal höher als an einer beliebigen Textstelle.[5] Auch zwischen nicht direkt aufeinanderfolgenden Buchstaben ist sicherlich mit Korrelationen zu rechnen. Ein simples Beispiel ist die Tatsache, dass zwei Zeichen nach einem `S` in einem deutschen Text die Wahrscheinlichkeit für ein `h` immer noch etwa 7.5 mal so hoch ist wie an einer beliebigen Stelle im Text.[5] Es wäre nun eine denkbare Annahme, dass solche Korrelationen nur über eine typische Skala λ existieren. Dies ist der Fall bei einem exponentiellen Abfall der Korrelationen: $C \sim e^{-x/\lambda}$, wobei x für den Abstand zweier Textstellen steht. Es gibt eine Reihe von Arbeiten aus den letzten 20 Jahren, die ein anderes Verhalten nahelegen (Schenkel et al., 1993; Amit et al., 1994; Ebeling und Pöschel, 1994; Ebeling und Neiman, 1995; Ebeling et al., 1995; Montemurro und Pury, 2002; Altmann et al., 2012). Dort werden vielfältige empirische Belege dafür vorgetragen, dass Korrelationen in Texten eher gemäß einer Potenzfunktion abfallen, also wie $C \sim \frac{1}{x^\alpha}$. Ein solches Ausklingen geht wesentlich langsamer vonstatten als im exponentiellen Fall. Im Gegensatz zur Exponentialfunktion gibt es bei einer Potenzfunktion auch keine typische Skala mehr, auf der die unterschiedlichen Teile eines Textes miteinander wechselwirken. Die graphische Darstellung in Abbildung 1.1 verdeutlicht dies. Der Graph der Funktion $y = \frac{1}{x^\alpha}$ hat unabhängig vom betrachteten Wertebereich dieselbe Form, während $y = e^{-\frac{x}{\lambda}}$ über kleine Bereiche $x \ll \lambda$ einer Geraden ähnelt und über große Bereiche $x \gg \lambda$ mehr und mehr an einen rechten Winkel in Form eines L erinnert.

Systeme, deren interne Korrelationen mit dem Abstand wie $\frac{1}{x^\alpha}$ abfallen, werden *skalenfrei* genannt, die entsprechenden Korrelationen werden als *long range correlations* bezeichnet. In der Physik ist skalenfreies Verhalten aus vielen Zusammenhängen bekannt, zum Beispiel aus der Thermodynamik der Phasenübergänge. Der Nachweis von *long range*

[5] ausgezählt an Bebel (2004a).

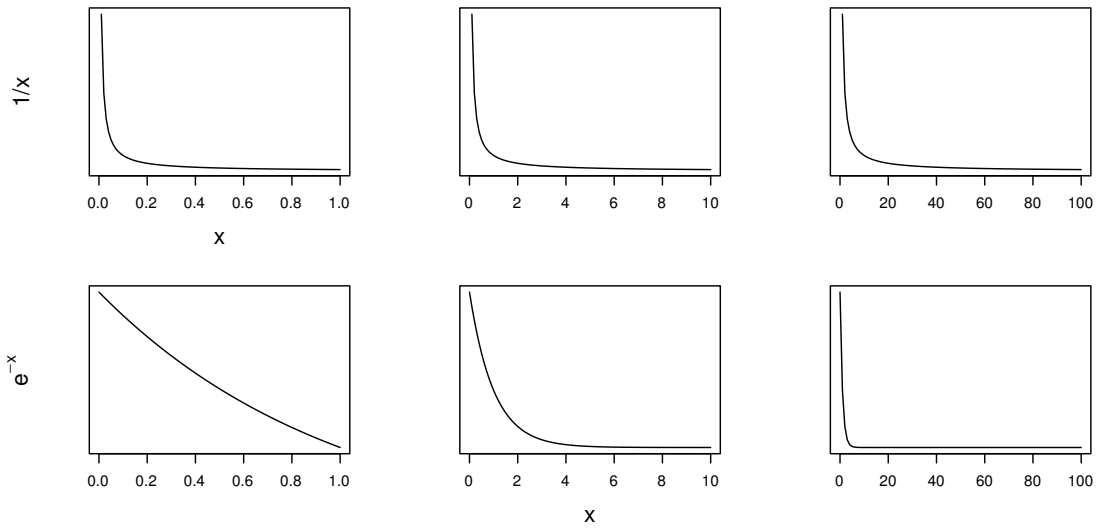


Abbildung 1.1: Die Potenzfunktion $y = \frac{1}{x}$ und eine exponentiell abfallende Funktion $y = e^{-x}$ in drei verschiedenen Wertebereichen. Links ist jeweils der Wertebereich von 0 bis 1 dargestellt, in der Mitte von 0 bis 10 und rechts von 0 bis 100. Die funktionale Form der Potenzfunktion $1/x$ ist jeweils identisch, während die Exponentialfunktion e^{-x} über die drei unterschiedlichen Wertebereiche einen sehr unterschiedlichen Anblick bietet. Im Gegensatz zur Exponentialfunktion, die eine eindeutige Skala λ kennt (hier ist $\lambda = 1$), verhält sich die Potenzfunktion skalenfrei.

correlations in natürlichsprachigen Texten gelingt den zitierten Arbeiten mit teilweise recht abstrakten und mathematisch anspruchsvollen Überlegungen und Untersuchungen. Abbildung 1.2 veranschaulicht ihr Vorhandensein anhand eines einfachen und anschaulichen Experimentes: Das Vorkommen oder Nichtvorkommen eines `i` an einer bestimmten Stelle in Bebel (2004a) korreliert mit dem Vorkommen oder Nichtvorkommen von `i` einige Zeichen später im Text. Die Stärke der Korrelation schwächt sich gemäß einer Potenzfunktion ab. Das Experiment alleine betrifft nur einen einzigen Buchstaben in einem einzigen Text und soll an dieser Stelle lediglich den Charakter der in den zitierten wesentlich umfänglicheren Studien gewonnenen Ergebnisse illustrieren. Neben diesen direkten empirischen Befunden motiviert aus linguistischer Sicht bereits die Tatsache, dass in menschlicher Sprache komplexe hierarchische Strukturen in eine lineare Aneinanderreihung von Elementen übertragen werden müssen, die Existenz langreichweitiger Korrelationen: Diese Struktur lässt Zusammenhänge zwischen relativ weit entfernten Elementen natürlich erscheinen.

Charakteristika skalenfreien Verhaltens von Texten zeigen sich aber auch in anders gelagerten Arbeiten. Golcher (2005, 2007b) untersucht die Gesamtzahl aller Wiederholungen in Texten bezogen auf die Textlänge. Es zeigt sich, dass die so definierte Wiederholungsrate in guter Näherung eine Konstante ist, wesentlich unabhängig von Sprache,

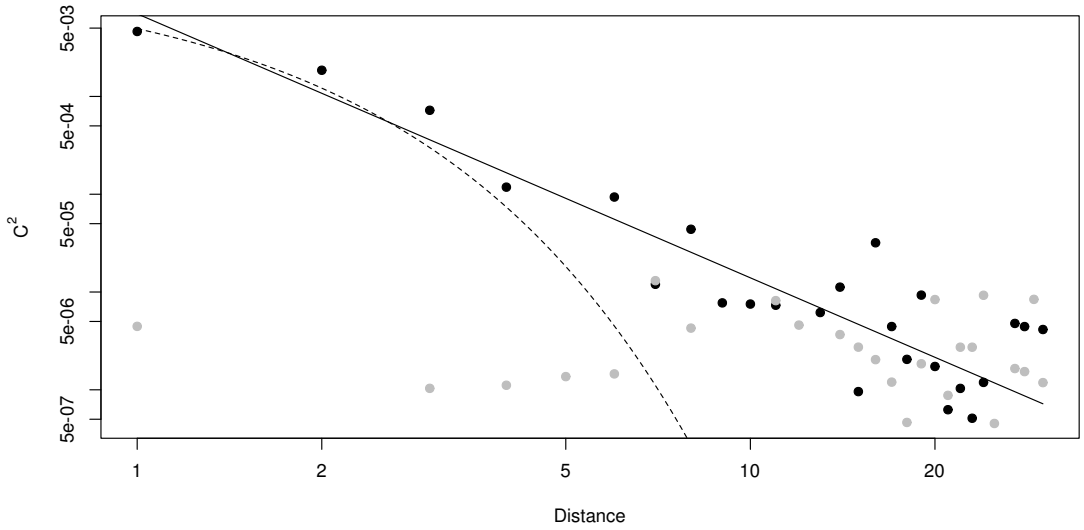


Abbildung 1.2: Charakteristischer Abfall von Korrelationen zwischen Buchstaben in Texten. Datengrundlage ist Bebel (2004a). Die Zeichenkette wird umgewandelt in eine Reihe v von 1 (wenn an dieser Stelle im Text ein `i` steht) und 0 (sonst). Die Zeichenkette `Dri_Chinisin` führt zum Beispiel zu $v = 001000101010$. Die X-Achse zeigt die Zahl der Zeichen zwischen zwei Textstellen, die Y-Achse die quadrierte Korrelation der Werte in v über diesen Abstand. Im Beispiel ergibt sich zum Beispiel für einen Abstand von 3 Zeichen eine Korrelation von $\hat{\varrho}(v_{3\ldots n}, v_{1\ldots n-3}) = \hat{\varrho}(000101010, 001000101) = -0.5$. Die Gerade entspricht $y = \frac{0.007}{Distance^{2.7}}$. Die grauen Punkte zeigen zum Vergleich denselben Text mit randomisierter Buchstabenreihenfolge, dh. mit rein zufälligen Korrelationen. Die gestrichelte Linie deutet einen möglichen exponentiellen Abfall an. Bis zu einem Abstand von 20 Zeichen kann man einen Abfall der Korrelationen entlang dem eingezeichneten Potenzgesetz erkennen. Dies ist das Merkmal langreichweitiger skalenfreier Korrelationen. Für noch größere Abstände geht die Kurve in das Rauschen zufälliger Korrelationen über. Andere Buchstaben zeigen ein ähnliches Verhalten. Die Korrelation wurde quadriert, um negativen und positiven Korrelationen dasselbe Vorzeichen zu geben.

Schriftsystem und auch von der Textlänge. Diese bemerkenswerte Konstanz ergibt sich nur, wenn alle Wiederholungen unabhängig von ihrer Länge gezählt werden. Die maximale Länge L_m der Wiederholungen nimmt mit der Textlänge n gemäß der Funktion $L_m \approx \frac{1}{2} n^{\frac{1}{3}}$ zu. Sie verhält sich also wiederum gemäß einer Potenzfunktion. Die Existenz einer festen Skala würde ein viel langsameres Wachstum gemäß $\log(\frac{n}{\lambda})$ nahelegen, wie das beispielsweise für eine zufällige Zeichenfolge der Fall ist. Man kann argumentieren, dass auch Zipf's Gesetz (Zipf, 1949) ein Beispiel für skalenfreies Verhalten ist. Es besagt,

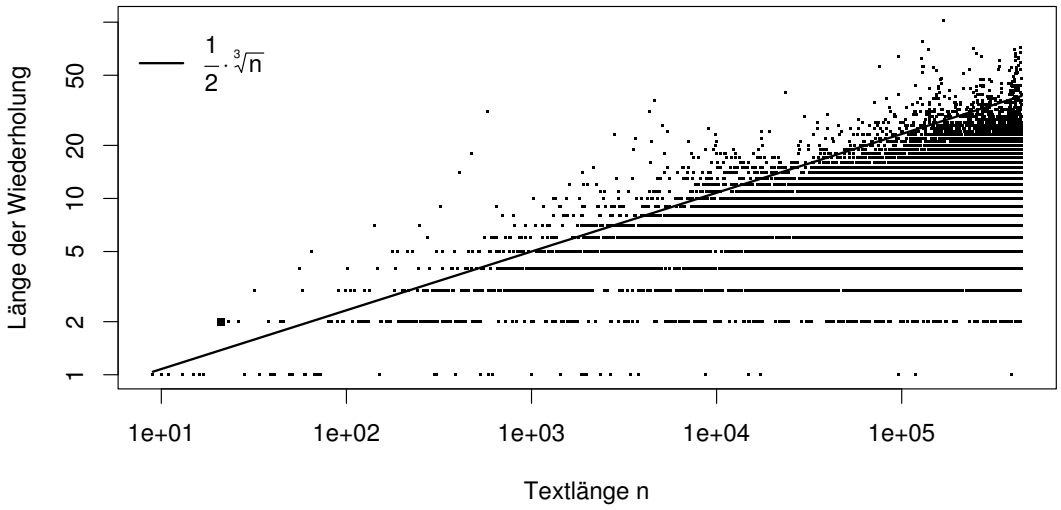


Abbildung 1.3: Die Länge der sich wiederholenden Zeichenketten in Bebel (2004a) aufgetragen über der Textlänge. Der Text beginnt mit den Worten `Aus meinem Leben. August Bebel`. Nach dem 24. Buchstaben (X-Achse) endet mit dem `g` eine Wiederholung (`Au`) der Länge 2 (Y-Achse). Diese Wiederholung erscheint links unten im Bild als vergrößerter Datenpunkt. Die durchgezogene Linie, die ungefähr der maximalen Länge L_{max} der Wiederholungen bei Textlänge n folgt, beschreibt die Funktion $L_{max} = \frac{1}{2}\sqrt[3]{n} = \frac{1}{2}n^{\frac{1}{3}}$.

dass in einer geordneten Frequenzliste die Häufigkeit von Wörtern in einem Text gemäß einem Potenzgesetz abnimmt.

Untersucht man nun nur kurze Zeichenketten oder allgemein nur kurze Folgen sprachlicher Elemente oder Ereignisse, bleiben diese langreichweitigen Wechselwirkungen notwendigerweise unberücksichtigt. Es ist eine bislang ungeklärte und sogar ungestellte Frage, was für Informationen in ihnen stecken und ob sich aus ihnen weitergehende theoretische und linguistische Erkenntnisse oder technische Anwendungen gewinnen lassen.

Die bisherige Forschung, die sich mit der Korrelationsstruktur von Texten auseinandersetzt, untersucht direkt die statistischen Eigenschaften von Sprache, um etwas über das erzeugende System, dh. die Sprache selbst zu lernen. Aber sowohl die Forschungen zur Wiederholungsrate in Texten (Golcher, 2005, 2007b) als auch die Untersuchungen zu *langreichweitigen Korrelationen* (Schenkel et al., 1993; Amit et al., 1994; Ebeling und Pöschel, 1994; Ebeling und Neiman, 1995; Ebeling et al., 1995; Montemurro und Pury, 2002; Altmann et al., 2012) haben einen recht abstrakten Charakter. So interessant die so gewonnenen Erkenntnisse möglicherweise sein können, so schwierig ist es auch, ihre tatsächliche Relevanz für die Entschlüsselung der Sprache einzuschätzen oder gar genauer zu formulieren, oder sie überhaupt mit den übrigen Gebieten der Linguistik in Verbindung zu setzen.

Eine alternative Möglichkeit, das Potenzial vollständiger Substringhäufigkeitsdaten zu untersuchen, ist ihre Wirksamkeit in Anwendungen zu überprüfen. Diesen Weg beschreite ich in meiner Arbeit. Es bieten sich Fragestellungen an, für die qualitative und womöglich quantitative Vergleichbarkeit mit veröffentlichten Forschungsansätzen und -ergebnissen herstellbar ist. Eine notwendige Voraussetzung hierfür sind Forschungsgebiete, auf denen bereits vergleichbare Daten verwendet wurden.

Mit der *Morphologischen Induktion* und der *Stilometrie* wende ich mich konsequenterweise zwei Gebieten zu, auf denen bereits intensive Forschungen auf der Grundlage von Zeichenketten unternommen wurden, wenn auch nicht anhand der hier untersuchten *vollständigen Substringfrequenzen*, sondern ausnahmslos an vergleichsweise kleinen Ausschnitten. Die sich daraus ergebenden Vergleichsmöglichkeiten machen überprüfbar, ob sich wirklich ein substantieller Vorteil aus dieser Ausweitung ergibt.

Es ist keine ganz und gar neue Idee, diese beiden Fragestellungen gemeinsam und mit ähnlichen Methoden zu untersuchen. Einen ähnlichen Weg geht zum Beispiel Teahan (2000), auch wenn sich die Daten, von denen er ausgeht und die Methoden, die er verwendet, stark von denen in dieser Arbeit unterscheiden. Interessanterweise zitiert Teahan (2000) eine empirische Beobachtung, die in scharfem Gegensatz zu den oben zitierten Untersuchungen zur Korrelationsstruktur in Texten steht.

> Experiments with English text show that [...] models with an upper bound of five characters in their context perform competitvely against other [...] models based on shorter or longer length contexts (Teahan, 2000, S. 948).

Damit wird eine Skala eingeführt, mit 5 Zeichen sogar eine ziemlich kurze. Implizit geschieht ähnliches in vielen Arbeiten aus beiden behandelten Themenkomplexen, aber selten wird ein solch genauer Zahlenwert explizit benannt und gerechtfertigt. Ein solches Vorgehen steht meinen Folgerungen aus dem Vorhandensein skalenfreier langreichweitiger Korrelationen entgegen. Es ist ein Ziel der vorliegenden Arbeit, mit weiteren Fakten den Widerspruch zu beleuchten, der zwischen den zitierten mathematischen Untersuchungen (Schenkel et al., 1993; Amit et al., 1994; Ebeling und Pöschel, 1994; Ebeling und Neiman, 1995; Ebeling et al., 1995; Montemurro und Pury, 2002; Altmann et al., 2012) zur Korrelationsstruktur von Texten auf der einen Seite und der verbreiteten Praxis in den angewandten Bereichen der Sprachanalyse und -verarbeitung auf der anderen Seite besteht. Dabei kann man vor allem deshalb auf neue Impulse hoffen, da sich diese Arbeit gerade nicht im mathematischen Kontext bewegt, sondern sich genau an den konkreten, anwendungsorientierten Fragestellungen orientiert, die auch Teahan (2000) bearbeitet. Daher können meine Argumente gegen das Einführen einer festen Skala direkter gegen die Argumente im Sinne von Teahan (2000) abgewogen werden.

Wenn auch das direkte Zusammentreffen der beiden Forschungsgebiete *Morphologische Induktion* und *Stilometrie* wie bei Teahan (2000) eher einen Einzelfall darstellt, so finden sich dennoch in beiden ähnliche Strömungen und Ideen wieder. Dies betrifft insbesondere kompressionbasierte Verfahren und Algorithmen. In der Forschung zur automatischen morphologischen Segmentierung gibt es den Gedanken, dass genau diejenige morphologische Zerlegung eines Textes optimal ist, die die knappste Beschreibung des

Textes ermöglicht. Auf dem Gebiet der Stilometrie dagegen wurden verschiedene Theorien vorgeschlagen, die davon ausgehen, dass zwei Texte genau dann ähnlich sind, wenn einer sich mit Hilfe des anderen möglichst weitgehend komprimieren lässt.

Auch in dieser Arbeit, die von keiner solchen Grundannahme ausgeht, finden sich Verbindungen zwischen den beiden Gebieten *Morphologische Induktion* und *Stilometrie*. Diese ruhen weniger in gemeinsamen Grundannahmen der vorgestellten Verfahren, als vielmehr in einer ähnlichen Struktur der Ergebnisse. In beiden Fällen tragen die selteneren, längeren Zeichenketten auf eine ähnliche Weise erheblich zum Erfolg der Methoden bei.

Das heißt, die vorgestellten Verfahren sind über ihre Tauglichkeit für potenzielle Anwendungen hinaus in der Lage, linguistisch relevante Fragen aufzuwerfen und Erkenntnisse zu befördern. Die Segmentierung von Texten in morphologische Einheiten ist eine sehr grundlegende Aufgabe. Hier ist die Dekodierung sprachlicher Struktur ein direktes Ziel. In Konsequenz kann man hoffen, dass Verfahren, die besser in der Lage sind, die morphologischen Einheiten eines Textes zu erkennen, auch die morphologische Struktur einer Sprache als Ganzes angemessener modellieren. Dann ist es möglich, aus dem Vergleich der Performanz von Verfahrensvarianten Rückschlüsse auf Eigenschaften des erzeugenden Systems selbst, der Sprache, zu ziehen. In einer umfangreichen Evaluation der Methode anhand dreier Sprachen (Deutsch, Englisch und Türkisch) wird in Abschnitt 2.6 eine derartige Untersuchung durchgeführt. Die allgemeinste ableitbare Schlussfolgerung besteht in der Beobachtung, dass in allen untersuchten Situationen Algorithmen auf Grundlage der *logarithmierten* Substringhäufigkeiten zu durchgängig besseren Ergebnissen führen als Algorithmen, die direkt die absoluten Zahlen verwenden. Der Logarithmus gibt den kleineren Häufigkeiten im Vergleich zu den großen mehr Gewicht.

Auch aus den Untersuchungen zur *Stilometrie*, die nicht direkt an der Entschlüsselung sprachlicher Strukturen interessiert ist, ergeben sich theorierelevante Fragen und Schlussfolgerungen. So untersuche ich die Frage, auf welcher Ebene der Sprache die effektivste Information über stilistische Unterschiede gefunden werden kann: Auf der Ebene der reinen Oberflächenwortformen, auf der Ebene der Wortarten (im Folgenden POS-Ebene genannt), oder auf der Ebene der Grundformen (Lemmata)? Auch im Rahmen der *stilometrischen* Untersuchungen wird wieder ein Vergleich verschiedener Algorithmusvarianten durchgeführt und wieder zeigt sich konsistent in verschiedenen Datensätzen, dass es von Vorteil ist, auch die längeren und selteneren Zeichenketten mit einzubeziehen.

Methodisch wird an manchen Stellen der Arbeit Neuland betreten, an anderen werden Grenzen verschoben. In den Abschnitten 2.6.2 und 2.6.3 werden allgemeine und verallgemeinerte gemischte lineare Modelle zur Evaluation der Performanz des Segmentierungsalgorithmus herangezogen. Nach meinem Kenntnisstand wurden diese Modelle bisher nicht in diesem Kontext verwendet. Der Vorteil ist eine erhebliche Erhöhung der Auflösung bei der Erfassung feinster Unterschiede zwischen den Verfahrensvarianten. In den Abschnitten 2.6.1, 3.6.2 und 3.6.4 zeigen sich die Probleme durch Wiederholungen verunreinigter Korpora in drei verschiedenen Facetten. Für dieses Problem werden Lösungsansätze wachsender Komplexität vorgestellt. Dabei helfen die hier verwendeten Frequenzinformationen sowohl beim Aufspüren als auch beim Beseitigen von Fehlern und Problemen in der Korpusaufbereitung. Im Rahmen der stilometrischen Untersuchungen

in Kapitel 3 wird neben den rein stilometrisch relevanten Fragestellungen auch das in diesem Zusammenhang oft unterschätzte Zusammenspiel der Einflussgrößen Stil, Texttopic und Textgenre untersucht.

1.2 Die Gewinnung und Aufbereitung der Daten

Die vollständige Frequenzinformation zu allen Substrings in natürlichsprachigen geschriebenen Texten sind sowohl Thema als auch Datengrundlage der vorliegenden Arbeit.

Bei der Berechnung derartiger Häufigkeiten von *Zeichenketten* stößt man schnell auf ein Problem. Bereits der ausgesprochen kurze Text, der nur aus dem Wort „Wald“ besteht, enthält die 10 verschiedenen Zeichenketten `Wald`, `Wal`, `Wa`, `W`, `ald`, `al`, `a`, `ld`, `l`, `d`. Allgemein ist die Zahl N der Teilzeichenketten in einem Text der Länge n gleich $n(n+1)/2$. D.h., nimmt ein Text auch nur $1MB$ ein, so benötigt die Liste der in ihm vorkommenden Zeichenketten $\frac{10^6(10^6+1)}{2}B = 500000500000B$ oder $500GB$ Speicherplatz. $1MB$ ist für einen einzelnen Text nicht ungewöhnlich lang, für ein ganzes Korpus ist es winzig.

Allein der enorme Platzbedarf hielt bisher wohl viele Forscher davon ab, alle diese Daten wirklich zu sammeln. Statt dessen werden wir in den entsprechenden Kapiteln zum Forschungsstand sehen, dass ein im Normalfall nur kleiner Ausschnitt betrachtet wird. Gewöhnlich sind es entweder die häufigsten Ketten, die zum Einsatz kommen, oder Ketten bis zu einer bestimmten Länge.

Ich verwende *Suffixbäume* um die vollständigen Frequenzinformation über sämtliche Substrings der untersuchten Texte zu speichern. Diese in Platz und Zeit ökonomische Indexstruktur ist in der Lage derartige Datenmengen handhabbar zu machen.

Unter einem Baum versteht man in Mathematik und Informatik, wie auch in der Linguistik, im Allgemeinen eine Struktur mit genau einem Wurzelknoten, von dem ausgehend sich einzelne Pfade immer weiter verzweigen, bis sie in Blättern enden. Gewöhnlich haben die Verbindungen von einem Knoten zum anderen (die „Kanten“ oder „Äste“) Beschriftungen.

Suffix (bzw. Präfix) bezeichnet im Kontext dieser Arbeit meist nicht die entsprechenden morphologischen Begriffe wie sie allgemein in der Linguistik verwendet werden, sondern ein beliebiges End- bzw Anfangsstück einer beliebigen Zeichenkette. In diesem Sinne ist `ird gut` ein Suffix des Satzes `Alles wird gut`.

Der Suffixbaum eines Textes ist eine Baumstruktur, in der jeder Pfad von der Wurzel zu einem beliebigen Blatt ein Suffix des Textes repräsentiert: Hängt man die Beschriftungen aller Kanten auf dem Weg von der Wurzel bis zum Blatt aneinander, so entsteht ein Suffix des eingelesenen Textes. Das Konzept des *Suffixbaums* wurde 1973 von Peter Weiner unter dem Namen *bi-tree* eingeführt (Weiner, 1973).

Betrachten wir den Text `abrakadabrax`. Die Liste der Suffixe dieses Textes im oben beschriebenen Sinne ist[6]

[6] Ich verzichte auf den leeren String, der streng genommen auch ein Suffix jeden Textes ist.

`abrakadabrax`	`dabrax`
`brakadabrax`	`abrax`
`rakadabrax`	`brax`
`akadabrax`	`rax`
`kadabrax`	`ax`
`adabrax`	`x`

Der gesuchte Suffixbaum ist derjenige Baum, der alle diese Zeichenketten als Pfade von der Wurzel zu einem Blatt enthält und nur diese. Er ist in Abbildung 1.4 dargestellt. So

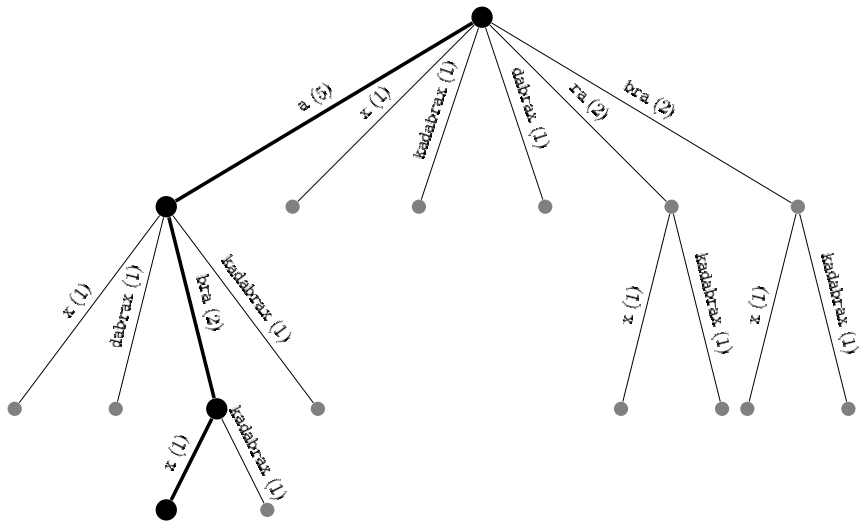


Abbildung 1.4: Der Suffixbaum des Textes `abrakadabrax`. Das Beispiel aus dem Text ist hervorgehoben. Die Zahlen in Klammern bezeichnen die Zahl der unter einem Knoten liegenden Blätter und damit zugleich die Zahl der Vorkommen der entsprechenden Zeichenkette. Folgt man einem Pfad im Baum von der Wurzel zu einem Blatt und hängt die Beschriftungen der Kanten aneinander, ergibt sich ein Suffix des Textes. Für den hervorgehobenen Pfad ergibt sich das Suffix `abrax` als `a+bra+x`.

ergibt sich das Suffix `abrax` als `a+bra+x`. Es ist in der Abbildung graphisch hervorgehoben.

Nicht nur jedes Suffix des Textes ist im Baum enthalten, sondern überhaupt jeder beliebige Teil des Textes, da jedes Teilstück des Gesamttextes auch Präfix eines Suffixes ist. Der einzige Unterschied ist, dass Suffixe an Blättern enden, bloße Teilstücke des Textes aber nicht.

Durch diesen Umstand ermöglichen Suffixbäume ein extrem schnelles Durchsuchen

großer Texte, da die Zeit, die benötigt wird, von der Wurzel her bis zu einer gewissen Tiefe in den Baum einzudringen, (weitgehend) unabhängig von der Gesamtgröße des Baumes ist. Daher ist auch die Zeit, die benötigt wird, um zu überprüfen, ob eine Zeichenkette im Text enthalten ist, unabhängig von der Länge des durchsuchten Textes. Sie ist lediglich proportional zur Länge der gesuchten Zeichenkette. Dieser Umstand ist für unsere Untersuchung zwar nicht direkt von Interesse, hat aber alle in dieser Arbeit vorgestellten Untersuchungen erheblich beschleunigt.

Da die Zahl der Substrings einer Zeichenkette quadratisch mit ihrer Länge steigt, ist es eine naheliegende Vermutung, dass der Aufwand für die Konstruktion eines Suffixbaumes dieser Zeichenkette ebenfalls quadratisch ansteigt.

Entgegen dieser Erwartung gab es von Beginn an Verfahren, Suffixbäume in linearer Zeit zu erstellen (Gusfield, 1997). Ich implementiere Ukkonens Algorithmus (Ukkonen, 1995), da er für die Anwendung in dieser Arbeit einen großen Vorteil aufweist. Die Zeichen des Textes werden gemäß ihrer Reihenfolge eines nach dem anderen in den Baum eingefügt. Dabei ist der entstehende Baum zu jedem Zeitpunkt ein vollständiger Suffixbaum des bisher eingelesenen Textes[7]. Dies ermöglicht es, an jeder Stelle des Textes auf die bisherige Häufigkeitsstatistik zuzugreifen.[8]

Obwohl der Kernalgorithmus von Ukkonen übernommen ist, war es für die verschiedenen Fragestellungen nötig, Modifikationen vorzunehmen, um die benötigten Informationen verwalten zu können.

Ergebnis des von Ukkonen entworfenen Algorithmus ist ein Suffixbaum, d.h. ein Baum, der genau alle Suffixe des eingelesenen Textes enthält. Für unsere Zwecke reicht dies noch nicht ganz aus. Zum einen muss nicht nur ermittelt werden, ob eine bestimmte Zeichenkette im Text enthalten ist oder nicht. Nötig ist die weitergehende Information wie oft sie vorkommt.

Daher habe ich die Struktur entsprechend erweitert, so dass diese Häufigkeitsinformation ebenfalls im Baum gespeichert wird. Ukkonens Algorithmus arbeitet so effektiv, dass nur lokale Änderungen am Baum vorgenommen werden. Das heißt, der Wurzelknoten wird nur selten besucht. Um die Information über die Gesamtzahl der vorkommenden Ketten aber im ganzen Baum zur Verfügung zu haben, wird die Nachricht über jeden erzeugten Knoten von der Stelle seiner Erzeugung bis zum Wurzelknoten durchgereicht. Gegebenenfalls können dabei sämtliche Informationen über sich wiederholende Zeichenketten ausgegeben werden. Da Knoten mit konstanter Rate erzeugt werden (Golcher, 2005, 2007b) und die Gesamttiefe des Baumes nur sehr langsam wächst, beeinflusst dies die Performanz des Verfahrens kaum.

Für den im Rahmen der *Morphologischen Segmentierung* von Texten (Kapitel 2) vorgestellten Algorithmus bedarf es einer weiteren Modifikation. Hier ist es nötig für den zu segmentierenden *Testtext* die Häufigkeiten aller seiner Substrings im *Trainingstext* zu gewinnen, und zwar ebenfalls in vertretbarer Zeit. Dazu wird erst aus dem

[7] Dies wird als die Online-Eigenschaft des Algorithmus bezeichnet.

[8] Eine eingehendere Beschreibung der Arbeitsweise von Ukkonens Algorithmus brächte an dieser Stelle keinen Mehrwert, da keine der technischen Einzelheiten im Kontext der untersuchten Fragestellung von Bedeutung ist. Ich verzichte daher auf eine eigene Darstellung und verweise direkt auf die Ausführungen in Gusfield (1997) bzw. in Ukkonen (1995) selbst.

Trainingstext wie üblich der Suffixbaum aufgebaut. Dann wird der Testtext eingelesen, *als ob* auch seine Substrings in den Baum aufgenommen werden sollten. Der Algorithmus ist allerdings so modifiziert, dass beim Einlesen des Testtextes, im Gegensatz zum Einlesen des Trainingstextes, alle Änderungen am Baum unterbleiben. Dies erlaubt es, die Trainings-Frequenzen der Substrings des Testtextes kompakt und geordnet auszugeben.

Darüber hinaus gilt es für die stilometrischen Untersuchungen in Kapitel 3, nicht nur einen, sondern zwei Texte zugleich in den Suffixbaum einzulesen, um die Frequenzen in beiden Texten vergleichen zu können. Es erweist sich als unproblematisch, den Algorithmus so zu erweitern, dass nicht nur einer, sondern zwei Texte zugleich in den Baum eingelesen werden können.

Nach dem einleitenden Abschnitt 1.1 und der detaillierteren Vorstellung der untersuchten Daten in diesem Abschnitt folgen nun in Kapitel 2 die Untersuchungen zur *Morphologischen Induktion*, bevor in Kapitel 3 die *Stilometrie* Thema sein wird.

2 Textsegmentierung mit partieller Strukturanalyse

2.1 Einleitung

Im folgenden Kapitel wird untersucht, inwieweit sich die in Abschnitt 1.2 beschriebenen *vollständigen Substringhäufigkeiten* eines Textes nutzen lassen, um die Einheiten, aus denen ein Text besteht, automatisch zu lernen. Diese vollständigen Substringfrequenzen bilden die einzige Datengrundlage. Explizites linguistisches Wissen findet keine Verwendung.[1]

Der Grundgedanke des Algorithmus ist einfach: Linguistisch bedeutsame Einheiten werden durch Zeichenketten repräsentiert, die als Ganzes häufiger zusammen auftreten und untereinander relativ frei kombiniert werden können. Dieser Gedanke ist nicht neu (2.4), neu ist die Konsequenz, mit der er umgesetzt wird, um Einheiten auf mehreren Ebenen zu identifizieren und die Vollständigkeit, mit der alle verfügbare Kontextinformation ausgenützt wird.

Aus praktischer Sicht ist die hier untersuchte Fragestellung von zentraler Bedeutung. Für viele computerlinguistische Anwendungen ist es wichtig, rohen Text vorab möglichst sauber in möglichst kleine sinntragende Einheiten zu zerlegen. Dies betrifft so verschiedene Fragestellungen wie die Erstellung von Wörterbüchern, den Bau von Suchmaschinen oder die Rechtschreibprüfung. Anders herum betrachtet gibt es nur wenige Anwendungen, für die solch eine Zerlegung *keine* notwendige Voraussetzung ist.[2]

Schon die automatische Zerlegung eines Textes in Wörter nach prinzipiell orthographischen Grundsätzen – die Tokenisierung – ist keine triviale Aufgabe. Sie ist natürlich noch ungleich schwieriger in Sprachen, die Wortgrenzen nicht explizit im Text markieren wie zum Beispiel viele asiatische Sprachen. Aber auch in denjenigen Sprachen, die das tun – z.B. alle europäischen Sprachen –, steht man vor einem großen Problem, wenn man eine automatische Zerlegung nicht nur in orthographische Wörter, sondern in sublexikalische, sinntragende Einheiten anstrebt. Vor allem in Sprachen mit hoch komplexer Morphologie wie Türkisch oder Finnisch ist dieses Problem offensichtlich, aber bereits das Deutsche mit seinen vielen Komposita und sogar das isolierende Englisch stellen für automatische Verfahren eine Herausforderung dar.

Dass die Frage praktische Relevanz hat, kann man schon daran ablesen, dass seit einigen Jahren ein regelmäßiger Wettbewerb organisiert wird, mit dem Ziel, die Forschung

[1] Drei marginale Ausnahmen werden in 2.3 auf Seite 25 f. diskutiert.

[2] Mein eigener Vorschlag für einen *stilometrischen* Algorithmus bildet eine Ausnahme, da er ja von denselben Daten ausgeht.

auf diesem Gebiet voranzubringen.[3]

Neben der praktischen Relevanz ist es eine fruchtbare theoretische Fragestellung, mit welchen algorithmischen Mitteln und mit welchen Daten man das Segmentierungsproblem mit welcher Qualität lösen kann. Dies ist nicht nur aus sich heraus eine interessante Frage, sondern kann auch Licht auf das Problem werfen, ob und bis zu welchem Grad menschliche Lerner aus Frequenzdaten allein die Struktur einer Sprache entschlüsseln können.[4] In Abschnitt 2.4 gehe ich auf ausgewählte Grundlagenforschung zum Thema ein.

Der Aufbau des Kapitels ist wie folgt: Abschnitt 2.2 gibt einen Abriss des theoretischen Hintergrunds, um der folgenden Diskussion und den empirischen Untersuchungen eine begriffliche Grundlage zu geben. Erst anschließend kann die gestellte Aufgabe präziser umrissen werden (Abschnitt 2.3). Es folgt mit Abschnitt 2.4 ein Überblick über das Forschungsfeld anhand der wichtigsten Arbeiten, die sich dieser und verwandten Aufgaben gestellt haben. Besondere Beachtung gilt hier den Ideen, die diesen Ansätzen zugrunde liegen. Anschließend wird der von mir entwickelte Algorithmus eingeführt (Abschnitt 2.5). Das Verfahren arbeitet konzeptuell zweistufig. Im ersten Schritt wird eine Menge möglicher Segmentierungen berechnet (Abschnitt 2.5.1). In einem zweiten, disambiguierenden Schritt wird aus dieser Menge eine bestimmte Segmentierung als die endgültige ausgewählt (2.5.2). An keiner Stelle gibt es freie numerische Parameter. Der zweite Schritt allerdings kombiniert verschiedene Bewertungsverfahren modular. Im Nachhinein erweist sich aber wiederum dieselbe Kombination in allen untersuchten Sprachen als optimal. Bestätigt sich dieses Ergebnis in einem weiteren Rahmen, ergibt sich ein sprachunabhängiges Segmentierungsverfahren.

Es folgt eine Darstellung verschiedener empirischer Untersuchungen (2.6). Die Datengrundlage bilden deutsche, englische und türkische Texte. Kurz zusammengefasst ist die Performanz des Algorithmus mindestens konkurrenzfähig im Vergleich zu veröffentlichten Alternativen. Wieso wirkliche Vergleichbarkeit derzeit nicht hergestellt werden kann ist im Rahmen der Vorstellung der aktuellen Forschungslandschaft Thema (2.4). Aus den Gemeinsamkeiten und Unterschieden zwischen den verschiedenen Sprachen lassen sich wertvolle Schlüsse ziehen. Auch die Abweichungen der berechneten Segmentierung von den Vorhersagen der Theorie und die Grenzen, an denen das Verfahren scheitern muss, ermöglichen linguistischen Erkenntnisgewinn.

2.2 Abriss der morphologischen Theorie und Notation

Morphologie bezeichnet das Studium der inneren Struktur von Wörtern und der Regeln, nach denen Wörter aus kleineren Einheiten gebildet werden. Da hier ein Verfahren untersucht wird, dass die Zerlegung von Texten in sublexikalische Einheiten anstrebt, wird es unumgänglich sein, einen Blick auf die relevanten theoretischen Begriffe zu werfen.

Im folgenden Abschnitt soll daher ein kurzer Abriss der morphologischen Terminologie

[3] `http://www.cis.hut.fi/morphochallenge2010/`; Wettbewerbe für Finnisch, Englisch, Deutsch, Arabisch und Türkisch.

[4] Vergleiche auch Goldsmith (2010, 13)[2]Hammarstroem2009a

gegeben werden. Nur mit solchem Rüstzeug können präzise Aussagen über den linguistischen Status der sowohl von anderen Forschern (2.4), als auch von meinem eigenen Algorithmus (2.5) vorgeschlagenen Strukturen getroffen werden. Ich werde mich weitgehend auf den Ausschnitt aus der Morphologie beschränken, der unmittelbar als Hintergrund für die empirischen Untersuchungen gebraucht wird.

Wesentliche Orientierungspunkte für das hier entworfene Definitionsgerüst sind die einführenden Aufsätze und Monographien Mugdan (1994) bzw. Lieber und Mugdan (2000), Grewendorf et al. (1987), Bauer (2003) und Wurzel (1984). Daneben habe ich mit Trost (2003) einschlägige Ausführungen aus dem computerlinguistischen Bereich und auch direkt Werke zum maschinellen Lernen von Morphologie (Tepper und Xia, 2010; Goldsmith, 2010) zu Rate gezogen. Auch die Ausführungen zu den theoretischen Grundlagen der in Abschnitt 2.4 referierten Arbeiten habe ich zur Kenntnis genommen und angemessen berücksichtigt.

Im folgenden werden also zentrale Begriffe der modernen Morphologie von Morph, Morphem und Allomorph bis hin zu Wort und Lexem eingeführt. Die meisten dieser morphologischen Einheiten sind im Strukturalismus verwurzelt. Grundlage dieser Denkströmung ist der Gedanke, dass die Bausteine eines Systems wie der Sprache keine Existenz unabhängig von diesem System haben können. Erst durch ihre kontrastive Abgrenzung von anderen Teilen des Systems kommen sie zu ihrer eigenständigen Existenz.

Nicht nur die theoretischen Begriffe, sondern auch die Algorithmen zum maschinellen Lernen von Morphologie wurzeln zu einem großen Teil im Strukturalismus, insbesondere in Werken von Zellig Harris aus den 1960'er Jahren. Darauf werde ich in 2.4 zurückkommen, wenn die bisherige Entwicklung und der aktuelle Stand des Forschungsgebietes vorgestellt werden.

Wie zum Beispiel Mugdan herausarbeitet, ist es bisher keinem morphologischen Formalismus gelungen, vollkommen befriedigende Definitionen für auch nur einen einzigen der von ihm diskutierten Begriffe zu finden. Welche Ansprüche an solch eine befriedigende Definition zu stellen wären, wird in Mugdans Text nicht explizit erläutert, aus den diskutierten Beispielen lässt sich aber schließen, dass sie mit dem intuitiven Urteil eines Linguisten übereinstimmen sollte. Eine ähnliche Situation findet sich in der übrigen oben zitierten Literatur.

Eine in diesem Zusammenhang recht plakative Stelle findet sich wiederum bei Mugdan (1994, 2546f): Er diskutiert den Begriff des Morphems, einen grundlegenden Terminus, der auch hier noch Thema sein wird. Ausgangspunkt ist eine kurzgefasste erste versuchsweise Definition („basic principle"). Anschließend wird ein konkretes Beispiel angeführt, das unter diese Definition fällt, was aber als ungünstig gesehen wird („seems rather strange"). Das Vorbild aber, im Vergleich mit dem dieses Urteil gefällt wird und den die Ausgangsdefinition folglich korrekt beschreiben sollte, bleibt im Dunkeln.

Mugdan äußert sich abschließend folgendermaßen zu diesem Problem:

> Linguistic analysis inevitably involves a certain amount of arbitrariness; decisions in borderline cases depend on the overall structure of the language, the purpose of the description (e.g., pedagogical considerations), or even individual taste.

Er begnügt sich konsequenterweise mit der Feststellung, dass jede Definition die Intuition eines Linguisten nur bis zu einem gewissen Grad abdecken kann. Er vermeidet entsprechend die Festlegung auf eine bestimmte Definition. Dies ist für einen einführenden theoretischen Text sicherlich angemessen.

Ich werde hier aber genau diesen Weg der festgelegten Definitionen gehen müssen, um eine eindeutige Diskussionsgrundlage zu schaffen. Die Fallstricke der einzelnen Begriffe werden kurz erläutert, um ihre jeweiligen Grenzen aufzuzeigen. Darüber hinaus akzeptiere natürlich auch ich die Tatsache, dass keine linguistische Definition genau die Fälle zu umfassen in der Lage ist, die dem Gefühl nach unter einem bestimmten Begriff subsumiert werden sollten. Schlimmer noch, auch meine Theorieskizze geht an ihren Rändern zwangsläufig in Begriffe über, die hier nicht erklärt werden, da der Platz begrenzt ist und sich diese Grenze der scharfen Begriffsbildung womöglich ohnehin nicht aufheben, sondern nur verschieben lässt.[5] Entsprechend beschränke ich mich auf eine für die angestrebte Diskussion ausreichend klare Festlegung der Kernbegriffe.

Ich beginne diese Skizze eines theoretischen Unterbaus mit der grundlegenden Definition des *Zeichens*. Ich vermeide die Bezeichnung „Buchstabe“, obwohl dies die ungefähre umgangssprachliche Entsprechung des Gemeinten wäre. „Buchstabe“ ist jedoch selbst für Alphabetsprachen ein wenig zu eingeschränkt, da schon die Leer- und Satzzeichen nicht enthalten wären. In Silbenschriften hingegen ist „Buchstabe“ völlig ungebräuchlich. Der verallgemeinernde Begriff *Schriftzeichen* hingegen wird meist als eingeschränkt auf Silbenschriften interpretiert. Um die folgende Definition mit der in dieser Arbeit so notwendigen wie natürlichen technischen Umsetzung in Übereinstimmung zu bringen und auch, um sie einfach zu halten, verwende ich eine technische Formulierung:

Definition 1 (Zeichen) *Ein Zeichen belegt im Unicodestandard*[6] *einen Codepunkt.*

Für Sprachen wie Deutsch und Englisch umfasst diese Definition natürlich auch die Menge der *Buchstaben*. Hinzu kommen allerdings alle Satzzeichen sowie das Leerzeichen. Wo das Gemeinte eindeutig ist, werde ich den Begriff „Buchstabe“ parallel verwenden.

Im vorliegenden Zusammenhang ist es sinnvoll, den Begriff der *Zeichenkette* bzw. des Strings zu definieren:

Definition 2 (Zeichenkette/String) *Eine Zeichenkette oder ein String ist eine Aneinanderreihung von Zeichen.*

Zeichenketten sind in `Schreibmaschinenschrift` gesetzt. Der besseren Sichtbarkeit wegen sind Leerzeichen meist durch Unterstriche repräsentiert.

Aus praktischen Gründen ist es zweckmäßig, zwischen *Zeichenkette* und *Text* zu unterscheiden:

Definition 3 (Text) *Ein Text ist eine Zeichenkette, die im jeweiligen Kontext nicht Teil einer längeren Zeichenkette ist.*

[5] „Habe ich die Begründungen erschöpft, so bin ich nun auf dem harten Felsen angelangt, und mein Spaten biegt sich zurück. Ich bin dann geneigt zu sagen: ‚So handle ich eben‘“ (Wittgenstein, 2001, §217)

[6] `http://www.unicode.org/`

Im Prinzip kann jede Zeichenkette als Teil einer längeren Zeichenkette gesehen werden. Wenn dies aus sachlichen Gründen nicht sinnvoll ist, sei es, dass es sich um ein ganzes Buch, oder einen Aufsatz handelt, sei es, dass ein Zufallstext in einer einzelnen Datei gemeint ist, spreche ich von einem *Text*. Von kurzen Beispieltexten abgesehen, handelt es sich schlicht um Texte im umgangssprachlichen Sinn.

Im Kontext dieser Arbeit ist der Begriff des *Korpus* im wesentlichen ein Synonym zu *Text*. Dies gilt für alle Stellen, an denen der im folgenden vorgestellte Algorithmus auf reale Daten angewandt wird. In diesem Fall kann „Korpus“ sowohl für einen einzelnen Text stehen (z.B. Bebel, 2004a), als auch für ein Korpus im linguistischen Sinne wie Francis und Kucera (1967). In diesem Fall allerdings wurden für die empirischen Studien dieser Arbeit oft alle oder einige Teiltexte des jeweiligen Korpus zu einem einzigen Text aneinandergehängt.

Es ist eine naheliegende Vermutung, dass ein Segmentierungsalgorithmus, der alleine auf der Basis von Häufigkeitsdaten arbeitet, auf unterster Ebene auch Zeichenkombinationen wie *sch* im Deutschen oder *sh* im Englischen als Segmente vorschlagen könnte. Obwohl später (2.5.1) dargelegt wird, warum Segmente dieser Art nicht in unserem Sinne sein können und wieso der hier vorgestellte Algorithmus sie auch von Anfang an vermeiden dürfte, führe ich hier dennoch die Begriffe *Digraph* und *Trigraph* ein:

Definition 4 (Di/Trigraph) *Ein Digraph (Trigraph) ist eine Zeichenkette der Länge zwei (drei), die ein einziges Graphem, repräsentiert.*

Als *Grapheme* werden die kleinsten bedeutungsunterscheidenden Einheiten des Schriftsystems einer Sprache verstanden, analog zu den *Phonemen* als der Menge ihrer kleinsten bedeutungsunterscheidenden lautlichen Einheiten. Da diese Arbeit ausschließlich auf der Grundlage geschriebener oder zumindest verschriftlichter Texte arbeitet, können wir Schnittstellen zur lautlichen Seite der Sprache auslassen.

Auf die naheliegende Frage, was wir hier unter *Bedeutung* verstehen wollen, werde ich weiter unten zurückkommen.

Ein wichtiger Schritt auf dem Weg zu einer morphologischen Terminologie ist die Definition des *sprachlichen Zeichens*. Im allgemeinen wird unter einem Zeichen im semiotischen Sinn ein Paar aus *Form* und *Inhalt* verstanden, für Details siehe zum Beispiel Mugdan (1994, S. 2543).

Definition 5 (sprachliches Zeichen) *Ein sprachliches Zeichen ist ein Paar aus einem Tupel M von Zeichenketten s_i mit $1 \leq i \leq n$ und $n \geq 1$ und einer sprachlichen Bedeutung. Die Elemente von M sind überschneidungsfreie Teilzeichenketten desselben Textes t, so dass $t = c^* s_1 c^+ s_2 c^+ \ldots, c^+ s_n c^*$. c^* steht hier für eine beliebige Zeichenkette beliebiger Länge, während c^+ für eine beliebige Zeichenkette der Länge ≥ 1 steht.*

Die Benennung *sprachliches Zeichen* vermeidet die Verwechslungsgefahr mit dem oben eingeführten Begriff des *Zeichens* im Sinne von „Buchstabe“ oder „Schriftzeichen“ (Definition 1).

Die Diskussion, was genau Bedeutung ist, halte ich an dieser Stelle noch einmal zurück.

Die Definition der Formseite eines sprachlichen Zeichens als Tupel erlaubt diskontinuierliche Konstituenten auf allen linguistischen Ebenen. So bilden die Zeichenketten s_1 = `Ich habe` und s_2 = `gelacht` in dem Text t_1 = `Ich habe schallend gelacht` als das Tupel $M_1 = (s_1, s_2)$ die erste Komponente eines sprachlichen Zeichens. Dasselbe gilt für den Text t_2 = `Gerenne` und das aus seinen Zeichenketten s_3 = `Ge` und s_4 = `e` gebildete Tupel $M_2 = (s_3, s_4)$. Ebenso sind gemäß Definition 5 auch die ein-elementigen Tupel (`schallend`) und (`renn`) oder auch die Beispieltexte selbst ((t_1) und (t_2)) erste Komponenten von sprachlichen Zeichen.

Erinnern wir uns nun daran, dass es das Ziel dieser Untersuchung sein soll, Texte in *linguistisch bedeutsame Einheiten* zu zerlegen (vgl. 2.1). Was genau solche Einheiten sein sollen war bisher noch unklar. Nun aber ist eine präzisere Festlegung möglich: Ein Text ist eine Zeichenkette (Definition 3). Jede Zerlegung wird wiederum aus Zeichenketten bestehen. Als einzige Kandidaten bieten sich nun genau jene Zeichenketten an, die allein oder zu mehreren die Formseite von *sprachlichen Zeichen* bilden. Da dies ein zentraler Begriff der gesamten Untersuchung sein wird, führe ich folgende Definition ein:

Definition 6 (sprachliches Segment) *Die Zeichenketten eines sprachlichen Zeichens heißen auch sprachliche Segmente. Ihre konkreten Vorkommen in einem bestimmten Text heißen Token sprachlicher Segmente.*

Dies wird zu unterscheiden sein von den Zeichenketten bzw. *Segmenten*, in die der Algorithmus einen Text zerlegt. D.h. einerseits wird von *sprachlichen Segmenten* die Rede sein, die – so nehmen wir es idealisierterweise an – der linguistischen Realität entsprechen. Andererseits werde ich von *Segmenten* sprechen, die Vorschläge des Algorithmus für *sprachliche Segmente* bezeichnen. In den allermeisten Situationen ist aus dem Kontext zweifelsfrei ersichtlich, was gemeint ist. Wo der Kontext jedoch nicht ausreicht, wird zwischen beiden Arten von Segmenten explizit unterschieden.

Was sind nun die kleinstmöglichen *sprachlichen Segmente*? Dazu muss geklärt sein wie sich *sprachliche Zeichen* zerlegen lassen:

Definition 7 (gültige Zerlegung) *Eine gültige Zerlegung eines sprachlichen Zeichens z ist jede Menge sprachlicher Zeichen $\{z_1, z_2 \ldots, z_m\}$, so dass*

1. *sich alle Zeichenketten der z_i genau zu den Zeichenketten von z verknüpfen lassen.*
2. *eine transparente Kombination der Bedeutungen der z_i zur Bedeutung von z existiert.*

Als *transparent* bezeichne ich eine für einen Sprecher nachvollziehbare Bedeutungskombination.

Im Sinne dieser Definition ließe sich ein sprachliches Zeichen z_1, dessen Zeichenkettentupel aus (`Gerenne`) besteht, in zwei Zeichen zerlegen deren Zeichenkettentupel das schon bekannte $M_2 = (s_3, s_4) = (\texttt{Ge}, \texttt{e})$ und $M_3 = (\texttt{renn})$ sind.

Diese Definition der *gültigen Zerlegung* ist notwendig, um *minimale sprachliche Zeichen* zu definieren:

Definition 8 (minimales sprachliches Zeichen / Morph) *Ein minimales sprachliches Zeichen ist ein sprachliches Zeichen, für das keine gültige Zerlegung existiert. Ein alternativer Begriff für minimales sprachliches Zeichen ist Morph.*

Vor einer Diskussion dieser Definition von *Morph* erweitere ich die Terminologie um einen davon abgeleiteten Begriff:

Definition 9 (minimales sprachliches Segment) *Die Zeichenketten eines minimalen sprachlichen Zeichens heißen minimale sprachliche Segmente. Ihre konkreten Vorkommen in einem bestimmten Text heißen Token minimaler sprachlicher Segmente.*

Definition 8 erklärt z.B. die Zeichenketten `ed` in `called` und `s` in `streets` zu zwei (Tupeln von Zeichenketten, die erste Komponenten jeweils eines) *minimalen sprachlichen Zeichen*(s sind). Dies ist Standard. Ebenso eindeutig ist `was` in `Ashley was boring` einzige Zeichenkette eines sprachlichen Zeichens, da es sich nicht weiter zerlegen lässt, ohne für sich genommen sinnlose Zeichenketten zu erhalten.

Definition 8 entscheidet sich aber auch dafür, Wörter[7] wie „Schornstein“ im Deutschen oder „cranberry“ im Englischen als *minimale sprachliche Zeichen* zu verstehen, da weder „Schorn“, noch „cran“ eine eigenständige Bedeutung zukommt. Derartige Einheiten werden als *Unikale* bezeichnet.

Ebenfalls wird in den Definitionen 8 bzw. 7 für den Fall eine Entscheidung getroffen, dass die Konstituenten eines zusammengesetzten Zeichens ihre Bedeutung nicht unverändert in das Gesamtzeichen einbringen. Mugdan (1994) nennt den Fall des englischen „blackberry“, dessen Konstituente „black“ nicht die Bedeutung des normalen Adjektivs „black“ habe, da Brombeeren auch grün sein können. Nichtsdestotrotz wird unter Muttersprachlern wohl Einigkeit darüber herrschen, dass das „black“ in „blackberry“ direkt mit dem gewöhnlichen Adjektiv „black“ zusammenhängt. Das heißt, die Verbindung von „black“ und „berry“ ist *transparent*, auch wenn es keine allgemein anwendbare Regel gibt, die aus den Bedeutungen dieser Einzelkomponenten die Bedeutung des Wortes „blackberry“ vorhersagen könnte.[8]

Auf derartige transparente Bedeutungskombinationen, welchen eine formseitige Verbindung zweier Zeichenketten entspricht, ist die parallele Definition der Kombination von Form und Bedeutung in Definition 7 ausgerichtet. Entsprechend wird „blackberry“ nach dieser Definition in zwei *minimale sprachliche Zeichen* zerlegt.

Der Rückgriff auf den Begriff der *Bedeutung* in der Definition des *sprachlichen Zeichens* (Definition 5) und damit auch des *Morphes* (Definition 8) ist die klassische Auffassung. Eine kleine Randbemerkung soll zeigen, dass sie nicht universell geteilt wird. Creutz und Lagus (2007, 3) schreiben: „Morfessor [ihr Algorithmus, meine Anmerkung] segments the input words into units called morphs. A lexicon of morphs is constructed where information about both the distributional nature (usage) and form of each morph is stored.“ Das heißt, ein *Morph* in diesem Sinne ist eine Zeichenkette zusammen mit Informationen

[7] Da hier nicht nur die reine Formseite gemeint ist, sind *Wörter* in Anführungszeichen gesetzt und nicht in `Schreibmaschinenschrift` wie Zeichenketten.

[8] Diese binäre Betrachtung des Begriffs *Transparenz* ist hier ausreichend. Für eine kontinuierliche Sichtweise vergleiche Wulff (2009)

über deren Verteilung. Was aber zeichnet derartige Zeichenketten aus? Dazu Creutz und Lagus (2007, 9): „We use the term lexicon to refer to an inventory of whatever information one might want to store regarding a set of morphs, including their interrelations." Somit bleibt der interessante Versuch, *Morph* ohne Rückgriff auf den Bedeutungsbegriff zu definieren, in einer Schleife gefangen.

In eine ähnliche Richtung weist die Skizze einer Definition von *word* von Goldwater et al. (2009, 22):

> [...] what constitutes a word: either a word is a unit that is statistically independent of other units, or it is a unit that helps to predict other units (but to a lesser degree than the beginning of a word predicts its end).

Diese rein statistische Definition kommt dem Grundgedanken, auf dem der hier dargestellte Algorithmus aufbaut – wie wir noch sehen werden – zwar recht nahe. Genau deshalb ist sie für unsere Zwecke vollkommen ungeeignet, denn hier soll ein theoretisch fundiertes Begriffsgebäude umrissen werden, das sich dann mit der automatischen Segmentierungsmethode vergleichen lässt. Eine rein operative Definition wäre zirkelhaft.

Daher bleibe ich bei der traditionellen Auffassung vom *Morph* als einem *sprachlichen Zeichen*, das sich nicht weiter in konstituierende *sprachliche Zeichen* zerlegen lässt wie sie in Definition 8 bereits präzisiert wurde.

Da diese Definition entscheidend auf dem Begriff der Bedeutung basiert, ist es an dieser Stelle nun notwendig, sich festzulegen, was hier unter dem Begriff *Bedeutung* verstanden werden soll. Da die hier verwendete Definition eng mit dem *Morph*-Begriff zusammenhängt wurde dieser wichtige Punkt so lange zurückgehalten.

Definition 10 (sprachliche Bedeutung) *Eine Menge an Eigenschaften von Tupeln von Zeichenketten,*

- *die nicht nur*
  - *die Zeichenketten selbst*
  - *oder den Text, in dem sie vorkommt,*

  betreffen
- *und die sich nicht nur auf ihre phonologische Qualität ihrer lautlichen Repräsentation beziehen*

heißt sprachliche Bedeutung. Die einzelnen Eigenschaften heißen Bedeutungskomponenten.

Diese Definition von *Bedeutung* bedarf einer Erklärung. Auf der einen Seite ist damit der Alltagsbegriff von Bedeutung eingeschlossen: Die Eigenschaft der Zeichenkette `Kuh`, ein weibliches Hausrind nach der ersten Kalbung zu bezeichnen, weist klar über die Zeichenkette selbst hinaus. Auf der anderen Seite steht die Zeichenkette `hfl` aus dem Text t_K = `Durch seine Kuhflecken-Optik ist er überall ein Blickfang`. Weder

ihre Eigenschaft, aus drei Buchstaben zu stehen, noch ihre Eigenschaft, in t_K vorzukommen, spricht ihr eine *sprachliche Bedeutung* zu. Ohne diese ist `hfl` folglich auch kein sprachliches Zeichen.

Definition 10 ist stark von der Morphemdefinition von (Grewendorf et al., 1987, 255) inspiriert:

> Ein Morphem ist die kleinste, in ihren verschiedenen Vorkommen als formal einheitlich identifizierbare Folge von Segmenten, der (wenigstens) eine als einheitlich identifizierbare außerphonologische Eigenschaft zugeordnet ist.

Grewendorf et al. selbst beziehen sich auf eine fast wortgleiche Definition von Wurzel.[9] Grewendorf et al., bzw. Wurzel erweitern den Bestandteil der *Bedeutung* der klassischen Morphemdefinitionen zur *außerphonologischen Eigenschaft*, um auch Fälle erfassen zu können, in der ein Morphem keine wie auch immer sprachexterne Bedeutung hat, sondern eine rein sprachinterne Funktion. Wurzel greift das Beispiel des *e* heraus, dass die femininen Nomina *Tante*, *Katze*, *Rose*, *Hose* und *Wonne* teilen. Dieses habe keine lexikalische Bedeutung, aber auch keine grammatische Bedeutung (oder Funktion), da es zum Beispiel im Singular wie im Plural gleichermaßen auftrete. Das Gemeinsame, die außerphonologische Eigenschaft, sei die Zugehörigkeit zur gleichen Flexionsklasse. Die Einbeziehung solcher Fälle wird sich als sehr passend für den hier vorgestellten Algorithmus zur morphologischen Segmentierung erweisen. Definition 10 übernimmt sie direkt.[10]

Ein Vergleich ergibt, dass die Definition von Grewendorf et al. bzw. von Wurzel sich recht gut deckt mit der hier gegebenen Definition 8 für *Morph* bzw. *minimales sprachliches Zeichen*. Grewendorf et al. (1987, 261) schreiben explizit, dass die Unterscheidung zwischen Morphem und Morph von ihnen nicht scharf gezogen wird.

Auch wenn diese Unterscheidung also vielleicht nicht unter allen Umständen nötig sein mag, führe ich sie hier doch ein. Mir scheint sie eine größere terminologische Präzision und Flexibilität zu ermöglichen und vielleicht auch eine größere Kohärenz mit verbreiteten Begriffssystemen herzustellen. Ich fahre also fort mit der Definition des *Morphems*:

Definition 11 (Morphem) *Ein Morphem ist eine Menge minimaler sprachlicher Zeichen, die folgende Bedingungen erfüllt:*

1. *Die Schnittmengen der Bedeutungskomponenten aller Elemente ist nicht leer.*
2. *Verschiedene Elemente sind komplementär verteilt, das heißt sie sind in keinem Kontext austauschbar.*

Die Elemente dieser Menge heißen Allomorphe des Morphems.

[9] „Ein Morphem ist die kleinste **vom Sprecher** in ihren verschiedenen Vorkommen als formal einheitlich identifizierbare Folge von Segmenten, der (wenigstens) eine als einheitlich identifizierbare außerphonologische Eigenschaft zugeordnet ist.“ (Wurzel, 1984, 38, Fettdruck von mir)

[10] Weder Grewendorf et al., noch Wurzel erweitern den Begriff der Bedeutung wie ich es hier tue. Statt dessen erweitern sie die Inhaltsseite des Morphems über die Bedeutung hinaus. Dies ist ein reiner Benennungsunterschied. Ich bin der Meinung, dass für unsere Diskussion ein griffiger Begriff wie *sprachliche Bedeutung* am zweckmäßigsten ist, auch wenn er ungewöhnlich weit gefasst ist.

Etwas unpräzise aber verständlicher formuliert: Ein Morphem ist eine Menge von Morphen, die dasselbe bedeuten und nie austauschbar sind.

Diese Definition ist im wesentlichen aus den in Mugdan (1994) und Bauer (2003) diskutierten Vorschlägen zusammengestellt. Da dort keine Definition gegeben wird, die in der Lage ist, alle konzeptuellen Probleme zu umgehen, sind sie auch mit der hiesigen Formulierung natürlich nicht vollständig beseitigt.

Nach Definition 11 gehören die zwei Formen des englischen unbestimmten Artikels „a“ und „an“ zum selben Morphem, da seine Ausprägung rein phonologisch bestimmt ist, und deshalb in komplementärer Verteilung auftritt. Ähnliches gilt für die meisten türkischen Flexionssuffixe, die wie das Pluralsuffix „ler/lar“ in zwei (oder vier) Varianten auftreten und deren Ausprägung der Vokalharmonie folgt.

Entsprechend werden die durch die Zeichenketten `ed`, `d`, `gave` und die *past-tense*-Formen anderer unregelmäßiger Verben repräsentierten *Morphe* zu einem Morphem[11] $m_1 = \{pasttense\}$ zusammengefasst, da ihnen Komponenten *sprachlicher Bedeutung* gemeinsam sind. Dies ist sehr ähnlich zur Auffassung in Bauer (2003, S. 19, Fußnote von mir):

> So *was* is a single morph that realises not only the lexeme[12] BE, (which is made up of a singel morpheme {be}) but also the morphemes {singular} and {past tense}.

Konsequenterweise würde das Morph `gave` neben seiner Rolle im Morphem m_1 ebenfalls zu einem aus `give`, `giv` und `gave` gebildeten Morphem $m_2 = \{give\}$ gehören.

Die Verben *begin*, *start* und *commence*[13] dagegen werden als zu unterschiedlichen Morphemen gehörig eingestuft. Sie erfüllen zwar die Bedingung der Bedeutungsüberschneidung, sind aber in vielen Kontexten gleichermaßen vorstellbar.

Im Rahmen dieser Arbeit sind auch lediglich durch Groß- und Kleinschreibung unterschiedene Ausformungen desselben Morphems (zB. `haus` im Nomen `das Haus` und im Verb `hausen`) als Allomorphe zu betrachten, da die Zeichenketten unterschiedlich sind. Damit können sie nicht Formseite desselben Morphs sein.

Die bisher dargelegte, rein konkatenative, dh. auf die aneinanderreihende Verknüpfung von Zeichenketten beschränkte, Beschreibung morphologischer Strukturen hat ihre Grenzen. So ist für viele deutsche Nomina die Pluralbildung nicht nur mit dem Anhängen einer Endung verbunden, sondern auch mit einer Veränderung des Stammes. Ein typisches Beispiel bildet das Singular-Plural-Paar „Haus“ und „Häuser“.

Natürlich gibt es angemessenere Beschreibungsrahmen für derartige Phänomene, wie zum Beispiel den als *item and process* (IP) bekannten Formalismus (Hockett, 1947; Mugdan, 1994, S. 2548f) oder den *word and paradigm*-Ansatz, wie ihn zum Beispiel Bauer (2003, 209ff) beschreibt.

Der von mir im folgenden Abschnitt (2.5) einzuführende Algorithmus allerdings legt sich implizit auf eine konkatenative Sichtweise der Morphologie fest und ist ohne tiefgreifende Erweiterungen nicht in der Lage, morphologische Prozesse auch tatsächlich als

[11] Morpheme werden durch hilfreiche Bezeichnungen in geschweiften Klammern notiert.
[12] was ein *Lexem* wiederum ist, sei an dieser Stelle ausgespart, s. aber Seite 23.
[13] Beispiel nach Mugdan (1994)

solche zu beschreiben. Im konkreten Beispiel hat das Verfahren keine Möglichkeit, den Umlaut in „Häuser“ aus dieser Wortform herauszulösen und möglicherweise sogar als zur Pluralendung gehörig zu erkennen. Entsprechend zweckmäßig ist es, bei einer konkatenativen Beschreibung zu bleiben und die damit verbundenen Probleme hinzunehmen.

Entsprechend wird die Wortform „Häuser“ hier nach den Definitionen 8 und 11 in zwei Zeichenketten `Häus` und `er` zerlegt, die die Formseiten zweier Morphe HÄUS[14] und ER darstellen. Das erste stellt ein Allomorph zu $\{haus\}$ dar, das zweite zeigt den Plural an.

Dieser ein wenig eingeschränkte Aufbau der morphologischen Theorie ist für Sprachen wie das Türkische durchaus angemessen, wo einzelne, klar voneinander abgrenzbare Suffixe mit eindeutiger und sehr modular aufgebauter Bedeutung wie zu einer Perlenkette (*string of pearls*) aneinandergereiht werden.

Wie wir gesehen haben ist eine rein konkatenative Betrachtungsweise für das Deutsche mit seinem Umlautsystem hingegen schon nur bedingt angemessen. In den semitischen Sprachen – wie dem Arabischen oder dem Hebräischen – aber kann von einer Aneinanderreihung keine Rede mehr sein. Lexikalisches Material und Flexionsmarkierungen müssen hier eher als miteinander verwoben beschrieben werden.

Dennoch stehe ich mit einer rein konkatenativ ausgerichteten Morphologie in der einschlägigen Forschung sicherlich nicht alleine. Ein willkürliches Beispiel ist Goldsmith (2010), der nur am Rande erwähnt, dass es nicht-konkatenative Morphologie gibt, um sich dann doch auf diese zu beschränken. Auch das obige Zitat aus Goldwater et al. (2009) (Seite 20) zeigt eine stark konkatenative Auffassung an. Eines der wenigen Gegenbeispiele ist Baroni et al. (2002).[15]

Es ist also bereits nicht einfach, den reinen Fachbegriff *Morphem* so zu definieren, dass er mit dem ausgebildeten Menschenverstand aller Linguisten in allen Fällen übereinstimmt. Ungleich schwieriger ist es aber, einen von sich aus noch komplexeren linguistischen Sachverhalt mit dem englischen Alltags*wort word* in Einklang zu bringen.

Mugdan (1994) zum Beispiel diskutiert die Definition von Wort als *Ortographical Unit*, *Phonological Unit*, *Lexical Unit* und *Grammatical Unit*. Er kommt jeweils zu dem Schluss, dass keine dieser Definitionen an den Alltagsbegriff anzupassen ist.

Entsprechend verzichte ich hier auf den Versuch einer wirklichen Formalisierung. Falls einmal doch vom Wort als linguistischem Begriff die Rede sein wird, ist darunter eine sehr oberflächennahe Interpretation des umgangssprachlichen Begriffs zu verstehen, im Sinne von *Wortform*. Im Rahmen dieser Arbeit wird sich dieser unscharfe Begriff als ausreichend erweisen.

Weil auch der Terminus *Lexem* weiter oben (Seite 22) bereits vorkam, sei er auch hier definiert. Oberflächlich betrachtet, ist er recht einfach zu fassen. In Anlehnung an Mugdan (1994, S. 2552) definiere ich:

[14] Wo hilfreich benenne ich konkrete Morphe durchgehend groß geschrieben.

[15] Die Autoren suchen genau dieses Problem zu überwinden. Sie nutzen orthographische Nähe und *mutual information* um zusammengehörige Paare wie „Haus“ und „Häus“ zu identifizieren. Allerdings gibt ihr Algorithmus keine Segmentierung zurück, sondern eine Zusammenordnung morphologisch zusammengehöriger Wortpaare. Eine knappe, aber zugängliche Zusammenfassung dieser Arbeit findet sich bei Roark und Sproat (2007).

Definition 12 *Ein Lexem ist eine Menge sprachlicher Zeichen, die in ihrer lexikalischen Bedeutung übereinstimmen, sich aber in ihrer grammatischen Funktion unterscheiden.*

Die Unterscheidung zwischen *lexikalischer* Bedeutung und *grammatischer* Funktion wäre natürlich eine eigene Diskussion wert.

Relativ zum *Lexem* lassen sich *Stamm* und *Affix* festlegen:

> In the simplest case, all forms of a lexeme can be derived from a single base by adding particular affixes ([...]). This base is traditionally called the stem. (Mugdan, 1994, S. 2553)

Da diese Begriffe hier nur am Rande verwendet werden, füge ich keine weitere Definition an, sondern beschränke mich auf dieses Zitat.

2.3 Charakterisierung und Einordnung der gestellten Aufgabe

Nach Klärung der notwendigen Terminologie kann die in der Einleitung zu diesem Kapitel recht informell und ein wenig implizit beschriebene Aufgabe präziser gefasst werden:

> **Ziel** des zu entwerfenden Algorithmus ist es, *natürlichsprachige Texte* in *Token* von *sprachlichen Segmenten* zu zerlegen.
>
> **Einziger Input** sind die in Kapitel 1.2 beschriebenen Daten.
>
> Es wird eine zumindest teilweise hierarchische Analyse angestrebt. Eine Annotation oder Klassifizierung der entstehenden Segmente ist dagegen kein Ziel.

Der Grundgedanke des Algorithmus kann als Hypothese formuliert werden und diese Hypothese zu überprüfen wird eines der zentralen Ziele dieser Untersuchung sein:

> Mithilfe der vollständigen Häufigkeitsstatistik aller Zeichenketten eines Textes kann man Segmente bestimmen, die die linguistische Definition des *sprachlichen Segmentes* (Definition 6) reproduzieren.

Damit im nächsten Abschnitt bisherige Arbeiten zum relevanten Themenkreis besprochen werden können, soll die dargestellte Fragestellung nun in die bestehende computerlinguistische Systematik eingeordnet werden.

Auf unterster Ebene werden Lösungsansätze der automatischen Sprachverarbeitung in zwei Kategorien eingeteilt: Regelbasierte Ansätze und maschinelle Lernverfahren.[16]

Regelbasierte Ansätze implementieren direkt sprachliches Wissen. Sie sind zwar ressourcenaufwendig und teuer, erreichen aber im Allgemeinen eine hohe Performanz. Um sie soll es hier nicht gehen, da das Interesse in dieser Arbeit in einer anderen Richtung liegt.

[16] Eine Einführung unter Bezugnahme auf morphologische Information in computerlinguistisch einsetzbaren Lexika ist Lüdeling (erscheint)

Maschinelle Lernverfahren verwenden Trainingsdaten um Strategien zur Lösung der Aufgabe zu entwickeln bzw. zu optimieren.

Es werden wiederum zwei Arten maschineller Lernverfahren unterschieden: *überwachte* und *unüberwachte* Ansätze.

Überwachte Ansätze haben zusätzlich zu den rohen Trainingsdaten die Möglichkeit, ihre eigene Performanz zu überprüfen und zu optimieren. Im kanonischen Fall bekommen sie Trainingsdaten, die mit Positivbeispielen versehen sind. Der Algorithmus passt seine Parameter an diesen *Goldstandard* an und überträgt ihn auf unbekannte Testdaten.

Unüberwachte Ansätze leiten mit Hilfe eines festen Algorithmus aus den rohen Trainingsdaten Parameterwerte ab.

Bei näherem Hinsehen ist die Unterscheidung zwischen *überwacht* und *unüberwacht* allerdings nicht ganz so klar zu ziehen: Ein *Goldstandard* positiver Beispiele ist nur eine Möglichkeit von vielen, ein System mit sprachlichem Wissen zu versorgen. Die oben erwähnten *regelbasierten* Verfahren sind hier natürlich das Extrem. Aber auch prinzipiell ist es wohl unmöglich, den Computer mit keinerlei Wissen über die Struktur des Problems zu versorgen und dennoch eine Lösung zu erwarten. Es ergibt sich ein Kontinuum von den *überwachten* zu den mehr und mehr *unüberwachten* Methoden.

Wenn die strikte Unterteilung in die Kategorien *überwacht* und *unüberwacht* aber nicht haltbar ist, was für eine Terminologie kann an ihre Stelle treten?

Es gibt mindestens zwei Möglichkeiten mit dem Problem umzugehen. So lockert Hammarström (2009, 14) die Definition von *unüberwacht*: „As little supervision, i.e., parameters, thresholds, human intervention, model selection during development etc., as possible“. Manning und Schütze (1999) dagegen plädieren dafür, die Frage anders zu stellen:[17] „*What knowledge sources are needed for use of this method?*“

Mein eigener Ansatz scheint schon alleine deswegen vollkommen unüberwacht zu sein,[18] da die einzigen übergebenen Daten aus unannotiertem Text gewonnene Frequenzdaten sind. Die vollständige Häufigkeitsverteilung aller in einem Trainingstext vorkommenden Zeichenketten scheint nun aber tatsächlich keinerlei sprachliches Wissen und auch keine Theorieabhängigkeit zu enthalten. Auf die bestehenden kleinen Ausnahmen sei kurz eingegangen.

In Definition 1 übernehme ich die Entscheidungen des Unicodekonsortiums darüber, was als ein *Zeichen* zählen soll. Dies ist für die hier empirisch untersuchten Sprachen Deutsch, Englisch und Türkisch keinerlei Problem. Es gibt aber durchaus Schriftsysteme, wo die Entscheidung darüber, was genau ein *Zeichen* ist, zu Diskussionen Anlass gibt, bzw. sprachliches Wissen kodiert. Ein Beispiel ist das indische Schriftsystem Devanagari, in dem die Vokale über ergänzende Zeichen oder Diakritika spezifiziert werden (Daniels und Bright, 1996). Aber auch dies sollte die Statistik sich wiederholender Zeichenketten kaum substantiell beeinflussen.

Es gibt noch zwei ähnliche kleine Abweichungen von der vollkommenen Unüberwachtheit. Erstens: Wie noch erläutert werden wird, führe ich Untersuchungen auch an Texten durch, die auf Kleinschreibung normiert wurden. Die Identifizierung von

[17] Kursivdruck im Original.

[18] Als solcher wird er auch von Hammarström (2009) hervorgehoben.

groß- und kleingeschriebenen Buchstaben ist eine Form sprachlichen Wissens.

Zweitens: Aus Gründen, die an gegebener Stelle näher diskutiert werden, billige ich dem Leerzeichen eine Sonderrolle zu. Etwas konkreter gesagt weiß der Algorithmus vorab, dass das Leerzeichen kein Zeichen im normalen Sinn ist, sondern lediglich ein Zwischenraum. Als Trennzeichen kann es keinem der begrenzten Elemente eindeutig zugewiesen werden. Auch dies kann als sprachliches Wissen betrachtet werden.

Es wird dem System aber weder mitgeteilt, dass Leerzeichen Wörter begrenzen, noch wird der Text vorab in Wörter zerlegt. Genau diese – viel ernstere – Form von sprachlichem Wissen geht in viele andere Verfahren ein: Das Leerzeichen als Worttrenner. Solche Verfahren sind aber schon infolgedessen kaum geeignet für Sprachen, die keine Wortzwischenräume schreiben. Für solche Schriftsysteme sind Tokenisierung und Morphologieerkennung zwei untrennbar verflochtene Aufgaben.

All diese Ausnahmen sind minimal. Der Kernalgorithmus, der den Text segmentiert, ist somit tatsächlich weitestgehend unüberwacht. Dieser Kernalgorithmus aber liefert noch keine eindeutigen Analysen, sondern berechnet eine Menge an mit gewissen Vorgaben kompatiblen Segmentierungen. Entsprechend ist ein Disambiguierungsschritt notwendig.

Inwiefern diese Disambiguierung überwacht verläuft und ob dies Konsequenzen für die Einsetzbarkeit des Verfahrens hat, wird an passender Stelle zu diskutieren sein (2.6.3).

2.4 Ideen und Arbeiten zur morphologischen Induktion: Ein Überblick

Wenden wir uns nun der Frage zu, was für Antworten auf die im vorherigen Abschnitt genauer charakterisierten Aufgabe bereits existieren. Da viele Autoren natürlich nicht genau das von mir gestellte Problem bearbeiten, aber dennoch relevante Ansätze vorstellen, ist es erforderlich, die Forschung zu einem etwas breiteren Gebiet zu referieren.

Das unüberwachte Lernen von Morphologie aus unannotiertem Text wird häufig mit dem Begriff *Morphologische Induktion*[19] (MI) bezeichnet,[20] auch wenn es eine große Vielfalt an alternativen Begriffen gibt.[21] Ich übernehme diese Terminologie in einem recht weiten Sinn:

Definition 13 (Morphologische Induktion) *Automatische Verfahren, deren (vielleicht nicht einziges) Ziel die Zerlegung von Texten oder orthographischen Wörtern in beliebige sublexikalische linguistische Einheiten ist, fallen unter den Begriff Morphologische Induktion, sofern sie weder einen Goldstandard benötigen, noch die Morphologie einer bestimmten Sprache direkt implementieren.*

[19] nach dem englischen *morphological induction*

[20] s. z.B. Roark und Sproat (2007, 5.1)

[21] „The problem is often referred to as Unsupervised Learning of Morphology, but also (Automatic) Induction of Morphology, Morpheme Discovery, Word Segmentation, Algorithmic Morphology, quantitative Morphsegmentierung (in German) and other variants have been used. “ (Hammarström, 2009, 1)

Ich spreche hier mit Überlegung nicht von *sprachlichen Segmenten* oder dergleichen wie es der von mir eingeführten Terminologie besser entspräche, sondern von *beliebigen sublexikalischen linguistischen Einheiten.* Viele Autoren führen entweder keine systematische Terminologie ein, und wenn doch, so decken sich ihre Definitionen untereinander natürlich nie vollständig. Daher gehe ich bei der Analyse fremder Ansätze, sofern das möglich ist, nicht weiter auf terminologische Probleme ein, als die Autoren selbst. Im Zweifelsfall übernehme ich deren Terminologie.

Es gibt Autoren, die zwischen der Zerlegung in Wörter (*Wortsegmentierung*) und der Zerlegung von Wörtern in kleinere Einheiten (*Morphemsegmentierung*) unterscheiden. Beispiele wären Mochihashi et al. (2009); Goldsmith (2010). In der Praxis sind die Fragen, die sich aus beiden Fragestellungen ergeben, und die vorgeschlagenen Lösungen so ähnlich, dass eine Trennung für die Beschreibung keinen sachlichen Vorteil bringt. Creutz und Lagus (2007, 4f) schreiben ähnlich:

> Unsupervised morphology induction is closely connected with the field of automatic word segmentation, that is, the segmentation of text without blanks into words (or sometimes morphemes).

Auch in der Realität der bereits existierenden Lösungsvorschläge macht es keinen Sinn, zwischen diesen beiden Facetten des Segmentierungsproblems zu unterscheiden. Ich werde mich hier dementsprechend auf beides beziehen.

Zu beachten ist allerdings, dass bei Wortsegmentierung notgedrungen mindestens ganze Sätze verarbeitet werden müssen. Morphemsegmentierung dagegen arbeitet oft mit Input auf Wortebene. Die Segmentierung längerer Texte oder Textstücke in *sprachliche Segmente* oder vergleichbare Entitäten dagegen ist in einer Minderheit der Arbeiten zu finden.

Nicht nur beim Input, oder bei der Natur der Outputsegmente, sondern auch bei der Struktur des Outputs sollten wir den Blick erweitern. Diese kann unter dem auf Seite 26 definierten MI-Begriff recht unterschiedlich sein. Drei Klassen können unterschieden werden: Erstens eine flache, eindimensionale Zerlegung des Inputs, so dass der Text eine Aneinanderreihung erkannter Segmente ist. Zweitens eine Zusammenordnung von erkannten Elementen zu Klassen, so dass Paradigmen oder Wortklassen entstehen. Drittens kann eine hierarchische Strukturierung angestrebt werden, mit dem Endziel der Erstellung von Analysebäumen.[22]

Alle drei Möglichkeiten werden realisiert. Die meisten Arbeiten gehen den ersten Weg und produzieren eine flache Zerlegung des Inputs, sei dieser nun ein Text, oder eine Wortliste. Es gibt aber auch Ansätze (z.B. Creutz und Lagus (2007); auch Yarowsky und Wicentowski (2000), Baroni (2003), Schone und Jurafsky (2001) und Goldsmith (2001) könnte man hier nennen.), die eine gleichzeitige Kategorisierung der Segmente in zumindest sehr grobe morphologische Klassen anstreben. Auch der dritte Weg wurde beschritten: So ist de Marcken (1996) in der Lage zusätzlich zur reinen Segmentierung eine hierarchische Ordnung zu liefern. Wie bereits angedeutet wird das auch die Struktur des von meinem Algorithmus ausgegebenen Ergebnisses sein.

[22] Eine ähnliche Aufteilung findet sich auch bei Hammarström (2009).

Bereits eine ungefähre Ordnung in die Vielfalt der Arbeiten zu MI zu bringen, ist keine leichte Aufgabe. Es gibt zwar eine Reihe durchaus brauchbarer Übersichten zum Thema. Als aktuell und informativ möchte ich vor allem Creutz und Lagus (2007); Goldsmith (2010); Hammarström (2009) und Roark und Sproat (2007) hervorheben. Bei Hammarström findet sich neben zahlreichen Artikeln zum Thema auch eine Liste mit weiteren Übersichtsartikeln. Allerdings schränkt er selbst ein (Hammarström, 2009, 15): „Nevertheless, there is no survey to date which is comprehensive and which discusses the ideas in the field critically." Auch ich habe keine wirklich vollständige Übersicht zum Thema gefunden und kann auch selbst keine in die Tiefe gehende Analyse der erschienenen Literatur und keinen wirklich vollständigen kritischen Überblick über die zugrundeliegenden Ideen und Verfahren geben. Bei ausreichender Tiefe wäre dies meines Erachtens tasächlich Stoff genug für eine eigene Arbeit. Stattdessen muss ich mich auf eine Skizze beschränken. In dieser Darstellung füge ich den unterschiedlichen, in den erwähnten Übersichtsartikeln gegebenen Klassifikationen der vorgeschlagenen Lösungen eine weitere hinzu. Es gibt keine so klar voneinander abgegrenzten Strömungen, dass eine Einteilung der Arbeiten sich von selbst anböte und keine der existierenden Strukturierungen erschien mir so fundiert, dass ich sie übernehmen wollte.

Ebenso wäre es selbstverständlich wünschenswert, die Performanz der vorgeschlagenen Ansätze untereinander quantitativ vergleichen zu können. Aber leider ist ein solcher Vergleich aus den bereits von Hammarström zusammengefassten Gründen unmöglich:

> We will not attempt a comparison in terms of accuracy figures as this is wholly impossible, not only because of the great variation in goals but also because most descriptions do not specify their algorithm(s) in enough detail. (Hammarström, 2009, 15)

Es ist in diesem Zusammenhang auch bedauerlich, dass es bis heute keine in der Community verbindliche Testsuite gibt, anhand derer sich die verschiedenen Ansätze vergleichen ließen. Ein Ansatz in dieser Richtung ist sicherlich das *Hutmegs evaluation package*[23], das es aber bisher ebenfalls nicht zu einer weitläufigen Verbreitung gebracht hat. Es wurde im Zusammenhang mit dem bereits angesprochene *Morphochallenge* entwickelt.

Ich unterscheide drei Gruppen von Aufsätzen, die großen Einfluss auf das Forschungsgebiet hatten und haben. Sie seien hier kurz vorgestellt.

Die Forschung zum unüberwachten Lernen von morphologischen Strukturen beginnt mit einem Werk des Strukturalismus, das in so gut wie jeder einschlägigen Arbeit erwähnt wird: Harris (1955).[24] Die Grundidee ist einfach: Elemente, die oft zusammen erscheinen, aber frei mit anderen Elementen kombiniert werden können, bilden eine Einheit. Konkret schlägt Harris vor, Morphemgrenzen genau dort zu setzen, wo die Zahl der möglichen Fortsetzungen eines Strings besonders groß ist.

Betrachten wir als Beispieltext August Bebels Autobiographie (Bebel, 2004b). Dort kommt die Zeichenkette `selbs` nur mit einer einzigen Fortsetzung vor: `selbst`. Die so

[23] `http://www.cis.hut.fi/projects/morpho/hutmegsdownloadform.shtml`

[24] Harris hat seine Ideen in späteren Arbeiten weiter ausgeführt (Harris, 1967, 1968).

verlängerte Zeichenkette dagegen hat 14 mögliche Fortsetzungen, darunter `b`, `e`, `g` und `k`. Diese Zunahme an Möglichkeiten kann man mit Harris (1955) als einen Hinweis darauf deuten, dass `selbst` eine linguistisch sinnvolle Zeichenkette darstellt. Harris prägte für die Zahl der existierenden Fortsetzungen einer Zeichenkette den Begriff *Successor variety*. Seit dieser frühen Begriffsbildung zieht sich der zugrundeliegende Gedanke durch einen großen Teil der Arbeiten zum Thema[25].

Ein weiterer unzweifelhaft sehr wichtiger Einfluss kommt von Shannons Arbeiten zur Datenübertragung (Shannon, 1948, 1951), die den Begriff der *Entropie* in die Informationstheorie überführten.

Unter Entropie wurde zu verschiedenen Zeiten in verschiedenen Wissenschaften unterschiedliches verstanden. Ursprünglich aus der Physik kommend spielt sie heute auch in allen Wissensbereichen, mit denen sich diese Arbeit auseinandersetzen möchte, eine nicht unbedeutende Rolle.

Es ist daher durchaus angebracht, den Ausprägungen und Veränderungen, die diese Größe auf ihrem Weg erfahren hat, eine kurze Diskussion zu widmen. Dies soll helfen, ihre Stellung und ihre Fundierung in Bezug auf die hier interessierenden Fragestellungen genauer einzuordnen.

Die *Entropie* begann ihre Geschichte um 1850 als eine Zustandsgröße der klassischen Thermodynamik. Dort half sie zu erklären, oder wenigstens zu beschreiben, warum manche Prozesse zwar spontan in der einen Richtung ablaufen, niemals aber ohne äußeres Zutun in der anderen: Die Entropie in einem isolierten System nimmt nie ab. Dies ist bekannt als der zweite Hauptsatz der Thermodynamik. Wie jede physikalische Größe hat die Entropie eine wohldefinierte Einheit $\left(\frac{Energie}{Temperatur} = \frac{J}{K}\right)$. Die augenscheinliche Gültigkeit des Zweiten Hauptsatzes und die Geschlossenheit der klassischen Thermodynamik rechtfertigt die Einführung der Entropie.

Im Rahmen der statistischen Thermodynamik gelang es Ludwig Boltzmann im Jahre 1877 eine Größe S zu formulieren, die die fundamentalen Eigenschaften der klassischen Entropie reproduziert und auf der Wahrscheinlichkeit der verschiedenen (Mikro-)Zustände aufbaut, die ein Systems haben kann.[26] Grob umrissen: Je wahrscheinlicher der (Makro-)Zustand des Systems, desto größer seine Entropie. Da das System seinen wahrscheinlichsten Zustand anstrebt, nimmt die Entropie nie von alleine ab.

Damit hat die Entropie den Sprung von einer phänomenologisch motivierten Größe zu einer fundamental abgeleiteten geschafft. Letzten Endes wird auch der zweite Hauptsatz

[25] Harris spricht wörtlich von *morpheme boundaries*. In der von mir eingeführten morphologischen Theorie macht das keinen Sinn, da ein *Morphem* eine Menge von *Morphen* ist und keine *Zeichenkette*. Wie oben auf Seite 26 f. diskutiert, werde ich über derartige Inkompatibilitäten zu der von mir verwendeten Terminologie soweit als möglich hinwegsehen.

[26] Nun ist die Entropie definiert als

$$S = -k_B \sum_i P_i \ln P_i$$

mit der Boltzmannkonstante $k_B = 1.38 \cdot 10^{-23} \frac{J}{K}$ und der Wahrscheinlichkeit P_i für den i-ten Zustand, den das System einnehmen kann. Gibt es nur einen Zustand mit $P_1 = 1$, so ist die Entropie $S = 0$, da $\ln(1) = 0$. In jeder anderen Situation ist sie größer und zwar umso größer, je mehr gleich (un)wahrscheinliche Zustände das System annehmen kann.

der Thermodynamik ableitbar und ist kein Postulat mehr.

In die Informationstheorie kam der Begriff als eine Neuformulierung bzw. Analogie des thermodynamischen Begriffs. Mathematisch ist der einzige Unterschied die fehlende Boltzmannkonstante, die nur physikalischen Sinn macht. Shannon nutzte den Begriff (unter anderem) dazu, die fundamentale Erkenntnis abzuleiten, dass ein Signal (eine *Zeichenkette*) nicht über eine Grenze hinaus verlustfrei komprimierbar ist. Diese Grenze ist durch die Entropie der Signalquelle vorgegeben.

Quantitativ war Shannons Formulierung revolutionär. Qualitativ ist die unterschiedliche Komprimierbarkeit in Abhängigkeit von der Entropie leicht zu verstehen. Aus der mathematischen Formulierung folgt schnell, dass eine Signalquelle, die immer dasselbe Signal aussendet, also absolut vorhersehbar ist, eine Entropie von 0 hat. Ein solches Signal kann demnach auf die Größe 0 komprimiert werden. Um zu wissen, dass heute die Sonne aufgegangen ist, braucht es keine Nachricht. Sie geht täglich auf. Eine entsprechende Nachricht hat eine Entropie von Null bzw. keinerlei Informationsgehalt. Im Gegensatz dazu ist ein Münzwurf überhaupt nicht vorhersagbar. Um das Ergebnis mitzuteilen, muss nach jedem Wurf eines von zwei Symbolen übertragen werden, zum Beispiel K für Kopf und Z für Zahl. Solch ein Signal lässt sich überhaupt nicht weiter komprimieren, seine Entropie ist maximal. Wird dagegen gewürfelt, ist es nicht nötig, jedes mal mitzuteilen, ob eine 6 gewürfelt wurde. Wesentlich kürzer ist es, immer nur anzugeben, wie viele Würfe bis zur nächsten 6 vergangen sind. Die Entropie ist weder minimal noch maximal.

Zusammengefasst: Ein vorhersagbares Signal, also ein Signal mit geringer Entropie, ist leicht zu komprimieren. Selbst in einer SMS hindern Abkürzungen die Lesbarkeit selten. Eine Telefonnummer dagegen muss immer ausgeschrieben werden: Ein unvorhersehbares Signal, bzw. eines mit hoher Entropie, erlaubt keine Komprimierung.

Shannon hat mit seiner Arbeit von 1948 die Informationstheorie begründet. In diesem Zusammenhang ist die Entropie schlicht eine mathematische Definition. Die Nützlichkeit dieser Definition begründet sich in der Bedeutung der von Shannon auf der Entropie aufgebauten Theoreme.

Es ist wesentlich, dass die Entropie der Informationstheorie nicht die Entropie der Physik ist. Dies kann man am einfachsten am Verlust der Einheit ablesen. Die Entropie der Physik hat eine aus Energie und Temperatur abgeleitete Einheit, während die Entropie der Informationstheorie keine Einheit hat. Die von Shannon eingeführte Einheit „Bit“ ist nur ein Name, den man dieser Zahl geben kann. Trotz der mathematischen Äquivalenz hat die Entropie nun auf eine ganz andere Abstraktionsstufe gewechselt. In der Physik diente sie dazu, Naturgesetze zu beschreiben und zu erklären. Bei Shannon erscheint sie als mathematische Entität in mathematischen und mathematisch bewiesenen Sätzen.

Nun kann man auch natürliche Sprache als ein Signal verstehen, für das man sich überlegen kann, bis zu welchem Maß seine Übertragung komprimierbar ist. Und tatsächlich hat bereits Shannon (1951) Experimente unternommen, um die Entropie der englischen Sprache abzuschätzen. Er beginnt seine Überlegungen mit der Vorhersagbarkeit des nächsten Buchstaben aus der Kenntnis der n vorhergehenden Buchstaben. Die Entropie der englischen Sprache ist dann definiert als der Grenzwert dieser Größe, wenn n gegen

unendlich geht. Dh, sie ist definiert über die Wahrscheinlichkeitsverteilung des nächsten Buchstaben, wenn man alle vorhergehenden Buchstaben kennt. Für n-Grammmodelle mit einem n in der Nähe von 1 gibt es wenig Schwierigkeiten.[27] Ob der Grenzwert für $n \to \infty$ allerdings existiert und eine realweltliche Deutung oder Bedeutung haben kann, scheint mir nicht gesichert. Hilberg (1990) hat in einem schlicht schönen Beitrag Shannons Daten neu interpretiert und äußert scharfe Bedenken an einer unteren Grenze der Entropie pro Buchstabe größer als Null.[28] Ebeling et al. (1995) wiederum widerspricht mit Verweis auf Levitin und Reingold (1994). Die Frage nach Existenz und Größe eines etwaigen Grenzwertes kann damit wohl als unentschieden gelten.

Was in der Praxis berechnet wird, ist nicht die tatsächliche Entropie einer hypothetischen Signalquelle an einer bestimmten Textstelle, sondern die Entropie eines einfachen Sprachmodells, dessen Parameter aus einer *Maximum Likelihood*-Schätzung (ML) gewonnen wurden. Allgemein steht ML für das Verfahren, dasjenige Modell auszuwählen, das die Wahrscheinlichkeit der tatsächlich beobachteten Daten maximiert. Wir werden auf dieses Modellauswahlprinzip im Laufe dieser Arbeit noch einige Male zurückkommen.

Für die Sprachmodelle, um die es hier geht, ist es gleichbedeutend mit der Schätzung von Wahrscheinlichkeiten aus relativen Häufigkeiten. Diese implizite Gleichsetzung von Häufigkeit und Wahrscheinlichkeit zieht sich durch die gesamte Literatur zum Thema, erscheint mir aber als durchaus gefährlich. Da diese Frage den konzeptuellen Kern des gesamten Wissensgebietes berührt, möchte ich diese Einschätzung ein wenig eingehender begründen.

Innerhalb der klassischen Physik ist der in der Entropie enthaltene Wahrscheinlichkeitsbegriff kein Problem und sehr gut beherrschbar. In Bezug auf das erzeugende System von Sprache scheint das nicht unbedingt gegeben.

Zumindest der frequentistischen Sicht nach sind Wahrscheinlichkeiten über den Grenzwert der relativen Häufigkeit mit der realen Welt verbunden. Die Schätzung von Wahrscheinlichkeiten aus Häufigkeiten setzt die Existenz stabiler Wahrscheinlichkeiten bzw. relativer Häufigkeiten voraus. Sonst existiert der Grenzwert nicht.

Dass die relativen Häufigkeiten auf der Wortebene schwanken, ist offensichtlich und keine Suchmaschine würde funktionieren, wenn es nicht so wäre: Die Häufigkeit von Wörtern wie „Anbieter“, „wetterfest“ und „zusammenknoten“ wird von Text zu Text unterschiedlich sein. Für Details zur Verteilung von Wörtern in Texten siehe auch Baayen (2001).

Diese Schwankungen der relativen Häufigkeiten stellt ihre Deutung als Wahrscheinlichkeiten in Frage.

Einige Einwände gegen diese Kritik können erwartet werden. Erstens kann man hoffen, dass Grenzwerte existieren, wenn die untersuchten Textmengen ausreichend groß sind. Konzeptuell wäre das so etwas wie der Mittelwert textspezifischer Wahrscheinlichkeiten. Ob wenigstens solche Mittelwerte existieren ist schwer zu belegen, aber auch schwer zu widerlegen. Diese Frage hängt eng mit den Begriffen der *Stationarität* und *Ergodizität*

[27] Aber siehe Seite 32.

[28] Er leitet einen Abfall mit der Wurzel der Textlänge ab. Pikant mutet seine sehr informiert wirkende Aussage, dass nach Shannon keine weiteren Daten zu diesem Thema mehr erhoben wurden.

von Sprache zusammen. Eine einführende Begriffserklärung und Diskussion findet sich bei Manning und Schütze (1999, 76).

Zwei weitere mögliche Argumentationslinien für die Gleichsetzung von Häufigkeiten und Wahrscheinlichkeiten im sprachlichen Kontext sind folgende: Zum einen kann man den Standpunkt einnehmen, dass die angesprochenen stabilen Grenzwerte innerhalb einzelner Texte durchaus existieren und daher die dazugehörigen Wahrscheinlichkeiten wenigstens auf Textniveau sinnhaft sind. Zum anderen kann man argumentieren, dass die Wahrscheinlichkeiten der nicht ganz so häufigen Inhaltswörter zwar stark schwanken, aber die Wahrscheinlichkeit der kürzeren und häufigeren Wörter oder Zeichenketten viel beständiger sind. Für diese ließen sich dann stabile Wahrscheinlichkeiten wenigstens als realistische Näherung annehmen, die über relative Häufigkeiten schätzbar wären.

Einen Überblick über das Thema und die damit verbundene Diskussion gibt Gries (2006, 2008). Hier soll ein kleines Experiment genügen um das Problem zu verdeutlichen: Verglichen wurden die zwei Bände von August Bebels Autobiographie (Bebel, 2004a,b). Beide Bände wurden vom selben Autor zum selben Thema in enger zeitlicher Folge verfasst. Wenn es irgendwo stabile Wahrscheinlichkeiten geben sollte, dann in einer derartigen Situation. Die Häufigkeiten der fünf häufigsten Bigramme[29] wurden bestimmt. Entgegen der Annahme ergibt ein Chiquadrattest eine signifikant unterschiedliche Verteilung in den zwei Texten ($\chi^2 = 17.0$, $df = 4$, $p = 0.0019$). Der geringe p-Wert ist natürlich auch eine Folge der hohen Zählungen (in allen Fällen > 27000), aber die relativen Verhältnisse schwanken auch für diese sehr häufigen Bigramme noch im niedrigen Prozentbereich, also verhältnismäßig stark. Kilgarriff (2005) führt ein Experiment ähnlichen Charakters durch und kommt zu entsprechenden quantitativen Ergebnissen.[30]

Meines Erachtens ist es nach dieser Diskussion zwar sinnvoll, interessant und erfolgversprechend, Häufigkeiten zu untersuchen und auf dieser Grundlage Anwendungen zu entwickeln. Diese Häufigkeiten als Wahrscheinlichkeiten zu deuten stellt aber viel zu oft nicht einmal eine gute Näherung dar, um von großem Nutzen zu sein.

Durch anscheinend theoretisch gut begründete Begriffe wie „Wahrscheinlichkeit“ und „Entropie“ wird leicht der Anschein einer theoretischen Herleitbarkeit von im Kern rein heuristischen Ansätzen erzeugt.

Möglicherweise mit inspiriert durch die von Shannon selbst hergestellte frühe und prominente Verbindung des sehr erfolgreichen Begriffs *Entropie* mit natürlicher Sprache wurde immer wieder versucht, ihn auch für konkrete computerlinguistische Probleme heranzuziehen.[31]

In diesem Kontext kann „Entropie“ meist als ein anderes Wort für „Unsicherheit“ gelesen werden. Bereits Harris' Begriff der *Successor variety* ist auch auf Grundlage dieses Entropiebegriffs formulierbar: Eine hohe Anzahl möglicher Fortsetzungen korrespondiert direkt mit einer großen Unsicherheit in Bezug auf den nächsten Buchstaben. Eine solche Verbindung zwischen den Ideen von Harris und Shannon konstatieren auch andere Autoren wie zum Beispiel Goldsmith (2010, S. 23).

[29] `ch`, `e_`, `en`, `er` und `n_`, wobei der Unterstrich das Leerzeichen repräsentiert

[30] Die Erwiderung von Gries (2005) halte ich in einem wesentlichen Punkt für fehlerhaft, der die Beurteilung der von ihm durchgeführten *post-hoc*-Tests betrifft.

[31] Goldsmith (2010, 23) vermutet sogar eine direkte Inspiration von Harris (1955) durch Shannon (1948).

Ein wichtiges Motiv für die Einbindung des Entropiebegriffs in die Computerlinguistik, und besonders in die Arbeiten zur *Morphologischen Induktion*, ist das Postulat, dass ein System genau dann Sprache besonders gut beschreibt, wenn es in der Lage ist, Sprache gut zu komprimieren. Eine Begründung für dieses Postulat oder eine empirische Rechtfertigung scheint nicht zu existieren. Mit ihm entsteht aber ganz natürlich eine Verbindung zur Entropie, da der Begriff der Entropie für die Theorie der Datenkomprimierung eine Schlüsselrolle spielt.

Eine sehr prominente Ausprägung dieser Denkrichtung werden wir mit den sogenannten MDL-Ansätzen kennen lernen (Abschnitt 2.4.2).

Von Anbeginn gründet sich damit ein großer Teil der Arbeiten zur *Morphologischen Induktion* konzeptuell auf Shannons Erkenntnisse und Begriffsbildungen und dies, obwohl diese Verwurzelung in keinem mir bekannten Text belastbar hergeleitet wird.

Die vorigen Absätze besprachen ausführlich die Bedeutung des Begriffes der Entropie im Allgemeinen und der Arbeiten von Shannon im Speziellen für das maschinelle Lernen von Morphologie.

Aus einer ganz anderen Richtung kommen die Arbeiten von Saffran und Kollegen (Saffran et al., 1996a,b) die das Feld ebenfalls stark beeinflusst haben. Hier werden experimentelle Untersuchungen präsentiert, die zeigen, dass menschliche Lerner Häufigkeiten in einem ansonsten unstrukturierten Sprachstrom nutzen können, um zusammengehörige Segmente zu identifizieren. Man kann diese Experimente als eine empirische Unterstützung von Harris' Ausgangsannahme lesen: Die Experimentatoren gaben ihren Versuchspersonen den Output eines einfachen Sprachmodells als Input und testeten, ob es ihnen gelingt, „Wörter“ zu identifizieren. Das Sprachmodell war derartig, dass eine Implementierung von Harris' Modell dieses Signal mit absoluter Präzision segmentiert hätte. Das gute Abschneiden der menschlichen Probanden kann als einen Hinweis darauf gesehen werden, dass das Gehirn einen ähnlichen Algorithmus implementiert. Dies wiederum ermutigt die automatisierte Verwendung solcher Verfahren. Die Arbeit wirkte also in zwei recht unterschiedliche Richtungen: Die technische und die psycholinguistische, vergleiche z.B. Goldwater et al. (2009).

Nach dieser Ausweitung der in 2.3 umrissenen Aufgabe auf die *Morphologische Induktion* insgesamt und der Erwähnung der theoretischen Einflüsse und empirischen Grundlagen folgt nun ein auszugsweiser Überblick über die wichtigsten und bekanntesten Arbeiten des Forschungsfeldes und über seine beherrschenden geistigen Strömungen.

2.4.1 Forschungstradition nach Harris

Da Zelig Harris' Ideen der erste konkrete gedankliche Anstoß in Richtung auf eine automatische Morphologieanalyse waren und man auch den von mir entwickelten Ansatz sachlich unter die davon inspirierten Arbeiten subsumieren kann, gehe ich zuerst etwas näher auf die diesbezügliche Forschungstradition ein.

Nach Harris' grundlegender Arbeit von 1955 dauerte es noch rund 20 Jahre, bis seine Ideen ernsthaft in lauffähige Algorithmen übersetzt wurden. Das erste Werk in einer

dann langen Reihe von Aufsätzen ist **Hafer und Weiss (1974)**.[32] Sie verwenden 4 Grundvariationen von Harris' Idee und kombinieren diese zu 15 Heuristiken, die sie miteinander vergleichen. Als Input dient eine aus dem Brownkorpus (Francis und Kucera, 1967) gewonnene Wortliste. Wie zu erwarten, finden wir also am Anfang eine Arbeit, die noch nicht so ambitioniert ist, ganze Texte zerlegen zu wollen, sondern sich mit einer Wortliste begnügt. Ziel ist eine Segmentierung dieser Wörter in *Stamm* und *Affix*.[33] Die Ergebnisse waren für die damalige Zeit durchaus beachtlich.

Bereits Hafer und Weiss scheitern („completely unsatisfactory") aber mit der naiven Implementierung von Harris' Idee an einem Problem, das zwar trivialer Herkunft ist, sich aber dennoch als ausgesprochen hartnäckig erwiesen hat. Da es zumindest im Hintergrund auch für die vorliegende Arbeit von Relevanz ist, möchte ich kurz etwas genauer darauf eingehen:

Die Autoren segmentieren in einem ersten Versuch jedes Wort genau an den Stellen, an denen die *successor variety* einen bestimmten Wert überschreitet. Dieses Verfahren zieht nicht in Betracht, dass die Zahl der Wörter, die ein Präfix teilen, mit der Länge dieses Präfixes abfällt. So gibt es in Bebel (2004a) 3320 Wörter, die mit `a` anfangen. Schon die häufigste Fortsetzung dieses Anfangsbuchstabens, das Präfix `au`, kommt nur noch auf 1249 Vorkommen. Dieser Abfall tritt nicht nur bei der Zahl der Vorkommen eines Präfixes auf, sondern auch bei der Zahl der *verschiedenen Fortsetzungen*, bzw. die *successor variety*. Das Präfix `a` hat 16 verschiedene Fortsetzungen, `au` dagegen nur noch 4 (`g`, `c`, `s` und `f`). Und dies, obwohl man, wenn ein `u` auf ein `a` folgt, im Normalfall wohl eher `au` als `a` als sprachlich relevante Zeichenkette sehen wollen würde. Dies gilt nicht nur für deutsche Texte und Wörter, die mit `a` anfangen, sondern ist ein allgemeines Phänomen.

Daher liefert Harris' Grundidee an sich nicht einmal eine grobe Näherung an eine linguistisch sinnvolle Segmentierung. Die Geschichte der Ansätze zur *Morphologischen Induktion* kann mit etwas Übertreibung gelesen werden als die Geschichte der Versuche, dieses Problem zu lösen oder zu umgehen. Dies gilt zumindest für die Klasse von Ansätzen, die sich explizit oder implizit auf Harris' Idee stützen.

Da ich sowohl hier bei der Darstellung der relevanten Literatur, als auch später bei der Darstellung meines eigenen Algorithmus (2.5) noch öfter auf dieses Problem zurückkommen werde, gebe ich ihm den Namen *Abfallproblem*.

Die Untersuchung von Hafer und Weiss bringt neben der Entdeckung dieses grundlegenden Stolpersteins einen gedanklichen Brückenschlag, der sich als für das ganze Forschungsfeld als sehr einflussreich erweisen sollte.

Eine der von Hafer und Weiss untersuchten Heuristiken basiert explizit auf der Shannon'schen Entropie[34] (Gleichung 26). Die Probleme, die ich in einem solchen Vorgehen sehe, wurden bereits etwas eingehender besprochen.

Ein praktischer Vorteil an der expliziten Einführung der Entropie ist, dass so die

[32] Überzeugend zusammengefasst in Goldsmith (2010).

[33] Auch dies wieder die (übersetzte) Originalterminologie: „[...] segmenting words into their stems and affixes" (Hafer und Weiss, 1974, 371)

[34] Wörtlich schreiben sie „[...] the entropy of the successor system for a test word prefix [...]" (Hafer und Weiss, 1974, S. 375)

Daten vollständiger ausgenutzt werden als in Harris' ursprünglicher Formulierung. Dort wird nur die Zahl möglicher Fortsetzungen berücksichtigt, während zur Berechnung der Entropie die Wahrscheinlichkeit bzw. die Häufigkeit dieser Fortsetzungen mit einfließen.

Außerdem stellt die Verwendung der Entropie eine mögliche Lösung des *Abfallproblems* dar, da mit der Berechnung der relativen Häufigkeit zur Schätzung der Wahrscheinlichkeit eine Normierung verbunden ist. Dass diese Lösung nicht wirklich tragfähig ist, lässt sich empirisch zeigen, siehe Abbildung 2.8 und die Diskussion dazu.

Ein praktischer Einwand allerdings gegen die Verwendung der Entropie für die Aufgabe der Segmentierung eines Textes wurde meines Wissens bisher nicht explizit formuliert: Die Berechnung der Entropie setzt die Summierung über eine Wahrscheinlichkeits*verteilung* voraus (vgl. Gleichung 26). Das bedeutet notwendigerweise, dass alle möglichen Fortsetzungen betrachtet werden, nicht nur die im jeweiligen Kontext vorkommenden. Dies wiederum macht es schwer, Grenzen von *sprachlichen Segmenten* je nach Kontext unterschiedlich zu setzen. Dies ist aber durchaus zu wünschen, da dieselbe Zeichenkette in Abhängigkeit von ihrer Umgebung sehr unterschiedliche Funktionen haben kann.

Ein ausgesprochen häufig zitiertes Werk, das sich direkt auf Harris (1955) beruft, ist **Dejean (1998)**.[35] Dejean hält sich im Ansatz ziemlich genau an die Vorgaben von Harris. In einem ersten Schritt ermittelt er mittels eines einfachen Cutoffs der *Successor Variety* Segmentgrenzen.[36] Nur die 100 häufigsten Morpheme gehen in die weitere Verarbeitung ein. Hierbei wird wiederum eine Anzahl von Heuristiken und Parametern eingesetzt.

Dejean (1998) ist, wie letztlich auch schon Hafer und Weiss (1974), ein typisches Beispiel dafür, wie leicht sprachliches Wissen auch in explizit als sprachunabhängig beschriebenen Ansätzen durch eine Hintertür doch wieder Zugang findet. Hier geschieht dies über die Annahme, dass Wörter aus genau einem Stamm und einem Suffix bestehen. Ähnliche Annahmen finden sich in vielen der zitierten Artikel, auch in den modernsten. So nehmen auch Mochihashi et al. (2009) noch einen bestimmten Wortaufbau an.

Auch wenn die Arbeit konzeptuell und methodisch hinter Hafer und Weiss (1974) zurückfällt, kommt Dejean doch das Verdienst zu, nachgewiesen zu haben, dass die Harris'sche Idee nicht nur für das Englische brauchbare Ergebnisse liefert, sondern für eine Vielzahl von Sprachen. Dejean untersucht neben Deutsch, Französisch und Englisch auch Swahili, Türkisch[37] und Vietnamesisch[38].

Das *Abfallproblem* geht Dejean nicht explizit an. Allerdings wählt er eine sehr hohe Segmentierungsschwelle für die *successor variety* und interessiert sich nur für das Auffinden der 100 häufigsten Morpheme. Es ist nicht unwahrscheinlich, dass dadurch das Problem

[35] Obwohl es ein wenig seltsam anmuten kann, dass eines der bekanntesten Werke dieser Forschungsrichtung ein Workshoppapier in fehlerbehaftetem Englisch ist.

[36] Originalterminologie: „[...] boundaries indicators between the elements which composed the sentences" (Dejean, 1998, 295)

[37] Hier macht er allerdings einen Fehler in der Beurteilung seiner Segmente. Nicht in jedem Fall sind die Morpheme, die er konstatiert, wirklich einzelne Morpheme (unabhängig von der Terminologie).

[38] Dejean konstatiert ohne jede Erklärung, dass seine Methode für das Vietnamesische keine Morpheme gefunden habe.

umgangen wird, da nur kurze und ähnlich lange Segmente mit einer entsprechend starken und gleichmäßigen Statistik betrachtet werden.

Schone und Jurafsky (2000, 2001) bieten ein sehr anschauliches Beispiel für einen Algorithmus voll durchdachter Heuristiken aus unterschiedlichen Bereichen mit etwa 5 verschiedenen numerischen Schwellwerten. Ziel ist zwar nur die Zerlegung von Wörtern, für diese Aufgabe wird aber das gesamte Trainingskorpus und der lineare Zusammenhang der Wörter im Text herangezogen.

Die Menge potentieller Affixe wird mit einem Harris-ähnlichen Algorithmus gewonnen: Als Affix zählt jede Zeichenkette, die ein Wort auf nicht eindeutige Weise fortsetzen kann, wie zum Beispiel das `en` in `leben`, da es noch andere Alternativen wie `st` gibt. Davon werden aber nur die Zeichenketten betrachtet, die ausreichend häufig vorkommen. Für die technische Realisierung werden Suffix*tries* verwendet, eine den Suffix*trees* sehr eng verwandte Indexstruktur. Im Gegensatz zu den in dieser Arbeit betrachteten Daten wird aber nur die Zahl der möglichen Fortsetzungen von Wortanfängen betrachtet, nicht die gesamten Substringhäufigkeiten des ganzen Trainingstextes.

Das sehr niedrigschwellige Kriterium zur Identifizierung von Affixen führt natürlich zu einer großen Menge an falsch Positiven. Die gewonnene Menge an Affixkandidaten wird daher mit einer rein distributiv verstandenen, berechenbaren Semantik verbunden und mit deren Hilfe weiter gefiltert. Das Verfahren ist als *Latent semantic analysis* bekannt, siehe zum Beispiel Manning und Schütze (1999, 554). Auf diesem Wege streben die Autoren eine Zusammenordnung der gefundenen Segmente an, zum Beispiel zu Flexionsparadigmen. Die Klassenbildung ist eine Eigenschaft, die diese Arbeit aus der Masse der vorgeschlagenen Algorithmen heraushebt. Eine prägnante Beschreibung findet sich in Creutz und Lagus (2002).

Ando und Lee (2003) zitieren Harris' Arbeiten zwar nicht, dennoch muss dieser Artikel in die Tradition seiner Idee gestellt werden (s. auch Creutz und Lagus (2007)). Analysiert werden japanische Texte. Diese Aufgabe ist ungleich anspruchsvoller als die Segmentierung von Texten europäischer Sprachen, da das Japanische wie viele asiatische Schriftsysteme keine Leerzeichen verwendet, um Wörter voneinander abzugrenzen.[39]

Dass sie damit notwendigerweise über die Segmentierung von Wörtern hinausgehen, rückt diese Arbeit in die Nähe meiner eigenen Untersuchungen, in denen es auch darum gehen soll, Sprache von der Satzebene ausgehend zu segmentieren.

Wie viele andere erweitern Ando und Lee (2003) Harris' Ansatz auf die schon aus der Diskussion von Hafer und Weiss's entropiebasiertem Ansatz bekannte Art (Seite 33), indem sie nicht nur die Zahl möglicher Fortsetzungen zählen, sondern auch deren jeweilige Häufigkeit berücksichtigen.

Die Grundidee ist eine Variante von Harris' Idee, die die Daten aus einem anderen Blickwinkel betrachtet als sonst üblich. Betrachtete Größe ist hier nicht die *successor variety*, also die Zahl der unterschiedlichen Fortsetzungen einer Zeichenkette. Vielmehr werden Stellen im Text ermittelt, an denen die angrenzenden Zeichenketten häufig sind, die diese Stelle umfassenden Ketten aber seltener.

Schiebt man zum Beispiel ein Fenster der Länge 3 über ein Vorkommen des Wortes

[39] Vergleiche auch die Diskussionen des Themas auf Seite 26 und 27.

`selbstverständlich` in Bebels Biographie (Bebel, 2004a), so bekommt man folgende Häufigkeitsverteilung, die einen Schnitt zwischen `selbst` und `verständlich` nahe legt:

```
bst          149
 stv          13
  tve         20
   ver      1038
```

Würde nur eine einzige Fensterbreite n verwendet, so würde das *Abfallproblem* hier nicht in Erscheinung treten, da nur Substrings gleicher Länge direkt miteinander verglichen würden.

Die Autoren beschränken sich aber nicht auf einen einzigen Wert von n, sondern ziehen Daten für verschiedene Werte dieses Parameters zu Rate. Hier macht es nun doch das *Abfallproblem* unmöglich, die entsprechenden Substringhäufigkeiten für verschiedene n direkt zu vergleichen. Ando und Lee nehmen Zuflucht zu einer nicht unkomplizierten Summierungsmethode und einer Mittelung über verschiedene Ordnungen von n.[40]

Obwohl Ando und Lee zwar die Möglichkeit von *suffix arrays*[41] zur Indexierung aller Substrings in einem Text erwähnen, beschränken sich dann aber doch auf n-Gramme endlicher Länge.

Die Arbeit von **Cohen et al. (2007)** ist ein aktuelleres Beispiel für ein recht ähnliches Vorgehen. Auch diese Autoren stellen sich die Aufgabe, Segmente in einem laufenden Text ohne markierte Wortzwischenräume aufzufinden. Der Text ist hier zwar nicht nur auf Chinesisch und (romanisiertem) Japanisch, sondern auch auf Englisch und Deutsch, welche ja die Wortgrenzen gewöhnlich explizit markieren. In diesem Fall wurden die Leerzeichen vorab entfernt. Segmentiert wird der Text einerseits wie schon bei Hafer und Weiss (1974) auf der Grundlage von *Entropie*-Maxima. Genauer gesagt wird die Entropie der Verteilung nach einem Kontext von n-Zeichen betrachtet.

Als zweiter Hinweis auf gültige Segmentgrenzen werden aber auch die Frequenzen der entstehenden Segmente selbst herangezogen. Das *Abfallproblem*, das daraus resultiert, dass kürzere Strings von vornherein häufiger sind, lösen die Autoren über die Normierung dieser Häufigkeiten: Die Daten werden z-transformiert, das heißt, statt der absoluten Frequenzen wird ihr Abstand vom Frequenzmittelwert für das jeweilige n ermittelt. Dieser Abstand wird in Einheiten der Standardabweichung umgerechnet. Als Beispiel führen sie die Zeichenketten `a` und `an` in englischen Texten an. Obwohl `a` als die kürzere Kette wesentlich häufiger ist, so ist die Häufigkeit von `an` doch wesentlich weiter entfernt von der mittleren Häufigkeit von Zeichenketten der Länge 2. Daher ist *an* in ihrer Analyse der bessere Kandidat für ein linguistisch sinnvolles Segment.

Das beschriebene Standardisierungsverfahren scheint zwar eine zumindest ungefähre Normalverteilung für n-Gramme anzunehmen, weil sonst die Normierung anhand der Standardabweichung fragwürdig wird. In Wirklichkeit sind n-Gramme natürlicher Texte

[40] Ihr Verfahren segmentiert an lokalen Maxima einer Funktion, die jeder Textposition einen Wert zuweist. Da lokale Maxima nicht direkt aneinandergrenzen können, benötigen sie einen numerischen Parameter t, der Ein-Zeichen-Worte ermöglicht.

[41] Wiederum eine eng mit *suffix trees* und *tries* verwandte Indexstruktur

Zipf-verteilt (Baroni, 2008). Eine solche Verteilung führt zu sehr wenigen sehr häufigen Elementen. Die Häufigkeit der übrigen Elemente fällt hyperbolisch ab. Solch eine Verteilung ist im Allgemeinen nur schwer beherrschbar und denkbar weit von einer Normalverteilung entfernt. Dennoch scheint die Performanz der Methode von Cohen et al. (2007) durch diese konzeptuelle Schwäche nicht übermäßig zu leiden.

Vom ebenfalls mit dem Entropiebegriff arbeitenden Ansatz von Hafer und Weiss (1974) unterscheiden sich Cohen et al. (2007) in drei wesentlichen Punkten: Erstens starten ihre Zeichenketten nicht nur am Wortanfang, sondern an einer beliebigen Stelle im Text. Zweitens ziehen sie nicht ausschließlich die Entropie an einer möglichen Segmentgrenze in Betracht, sondern auch die Frequenz der Segmente. Drittens z-skalieren sie nicht nur diese Frequenzen, sondern auch die Entropiewerte.

Feng et al. (2004) widmen sich wie Ando und Lee (2003) der Segmentierung von asiatischen Texten ohne Leerzeichen. In ihrem Fall handelt es sich nicht um Japanisch, sondern um Chinesisch.

Sie erweitern den Harris'schen Begriff der *successor variety* in einem wichtigen Punkt. Sie betrachten nicht nur die Zahl der unterschiedlichen Fortsetzungen (in ihrer Notation $R_{av}(s)$) eines Strings s, sondern auch die Zahl der unterschiedlichem *Zeichen*, die ihr vorausgehen können ($L_{av}(s)$). Ihr Segmentierungsalgorithmus baut dann auf dem Minimum dieser beiden Zahlen auf, die sie *accessor variety* $AV(s)$ nennen.

Wie für chinesische Texte sinnvoll, beschränken sie die maximale Länge der Segmente auf 6 *Zeichen*. Aber auch dies ist eine Form sprachlichen Wissens und müsste für andere Sprachen zumindest entsprechend angepasst werden. Nebenbei gesagt verwenden Varianten von Harris' Idee als Datengrundlage immer nur Substrings des Textes mit einer vorher festgesetzten maximalen Länge n, zumindest ist mir keine Ausnahme bekannt.

Jedem der so entstehenden Segmente wird über eine Gütefunktion ein Wert aufgrund von Segmentlänge und *accessor variety* zugeordnet. Die Segmentierung mit dem höchsten Gesamtwert wird schließlich ausgewählt. Die Autoren untersuchen verschiedene Gütefunktionen.

Da der Gedankengang dieses Rankings partielle Ähnlichkeit zum von mir erarbeiteten Disambiguierungskonzept (2.5.2) aufweist, werde ich bei dessen Beschreibung noch einmal auf diesen Artikel zurückkommen.

Das *Abfallproblem* wird von Feng et al. (2004) über die explizite Aufnahme der Wortlänge in die Gütefunktion angegangen.

2.4.2 Die bayesianischen Arbeiten

Die MI-Arbeiten, die in der Harris'schen Tradition stehen, könnte man als heuristisch bezeichnen. Ausgehend von der einen oder anderen Formulierung seiner im Kern strukturalistischen Idee werden Algorithmen implementiert, meist ohne eine weitere theoretische Fundierung. Eine scheinbare Ausnahme sind Rückgriffe auf den Entropiebegriff wie wir ihn unter anderem schon bei Hafer und Weiss (1974) gesehen haben. Es wurde aber bereits besprochen, dass diese meist implizite Argumentation auf der Übernahme eines in Physik und Informationstheorie recht mächtigen Begriffes in die Linguistik beruht. Die Nützlichkeit dieser Übernahme und Wohldefiniertheit des Begriffs in der Linguistik

wird aber kaum thematisiert.

Nicht ganz von den Arbeiten nach Harris zu trennen, aber doch einem anderen Geist verpflichtet ist ein Forschungsstrang, den ich als *bayesianisch* bezeichnen möchte. Folgendes ungefähre Vorgehen ist allen derartigen Verfahren gemein:

Jede konkurrierende Zerlegung des Textes wird mit einem Sprachmodell assoziiert. Creutz und Lagus (2007) umreißen das Wesen eines solchen Sprachmodells mit folgenden Worten: „The model of language (M) consists of a morph vocabulary, or a lexicon of morphs, and a grammar“. Aus der bedingten Wahrscheinlichkeit $P(T|M)$, mit der dieses Sprachmodell M den ursprünglichen Text T produziert, und aus der *a priori*-Wahrscheinlichkeit $P(M)$ für das Modell wird mittels des Bayesschen Gesetzes auf die *a posteriori*-Wahrscheinlichkeit $P(M|T)$ des Modells zurückgeschlossen:

$$P(M|T) = \frac{P(T|M)P(M)}{P(T)} \tag{2.1}$$

Diese Beziehung ist von universeller Bedeutung und kann verwendet werden um aus gemessenen Daten auf die Wahrscheinlichkeit eines Modells zurückzuschließen. Zweck dieser Berechnung ist meist, das bei Kenntnis der Daten wahrscheinlichste Modell auszuwählen. Im vorliegenden Kontext sind die gemessenen Daten der zu analysierende Text. Da die *a priori*-Wahrscheinlichkeit $P(T)$ des Textes T eine Konstante ist, kann man sie für Optimierungsrechnungen im Allgemeinen vernachlässigen. Die verwendeten Modelle sind gewöhnlich so einfach gehalten, dass sich $P(T|M)$ aus dem Modell berechnen lässt. Annahmen über die *a priori*-Wahrscheinlichkeit $P(M)$ des Modells tragen leicht einen gewissen ad-hoc-Charakter. Je nachdem, welcher *Prior* für das Modell gewählt wird, ergeben sich recht unterschiedliche Verfahren.

Allen derartigen Verfahren gemein ist aber die Notwendigkeit eines Suchalgorithmus. Der Raum der möglichen Modelle bzw. ihrer Parameter ist normalerweise vieldimensional und unendlich. Das beste oder wenigstens ein sehr gutes Modell auszuwählen ist häufig kein einfach zu lösendes Optimierungsproblem. Vor allem bei den modernsten Verfahren liegt großes Gewicht auf der Suche nach einem möglichst effizienten Suchverfahren (s. z.B. Goldwater et al. (2009)).

Ein Ansatz der bayesianischen Richtung ist **Snover et al. (2002); Snover und Brent (2001)**.[42] Das Ziel dieser Autoren ist die Zerlegung von Wortlisten in jeweils einen Stamm und ein Suffix. Für die analytischen (und/oder isolierenden) Sprachen, die sie untersuchen – Französisch, Englisch und Polnisch, – ist diese vereinfachte Sicht weitgehend wohl auch angemessen. Im engeren Sinne „language independent“ wie die Autoren in Snover et al. (2002) schreiben ist dies allerdings nicht. So folgen zum Beispiel die Morphologien agglutinierender Sprachen wie des Türkischen nicht diesem Muster.

Ein tieferes Problem der beiden Arbeiten scheint mir allerdings in ihrer Argumentation zu liegen, dass nicht nur der Nenner in Gleichung 2.1 konstant sei, sondern auch die bedingte Wahrscheinlichkeit der Daten $P(T|M)$ als konstant gleich 1 gesetzt werden könne. Ihre Begründung betont, dass der Text ja bereits feststehe und seine Wahrscheinlichkeit unter dem Modell somit gar keine Rolle spiele. Dies widerspricht aber klar der

[42] Eine kurze, zugängliche Zusammenfassung findet sich bei Roark und Sproat (2007).

Definition von $P(T|M)$ als der Wahrscheinlichkeit, dass T produziert wird, wenn M die Quelle ist. Diese kann nur 1 sein, wenn das Modell auch in Zukunft mit Sicherheit keinen anderen Text produzieren würde. Dies wäre kein Sprachmodell im intendierten Sinn. Somit würde der Sinn des Bayesschen Theorems verkannt.

Praktische Auswirkung dieser Operation ist, dass mit der Gleichung

$$P(M|T) = \frac{P(M)}{P(T)}$$

weitergerechnet wird, im Grunde sogar mit $P(M|T) = P(M)$, da $P(T)$ ja (tatsächlich) als Konstante gesehen werden kann. Snover et al. (2002) schreiben das genau so: „[...] the probability of the hypothesis given the data reduces to $Pr(Hyp)/c$“.

Abstrahiert man von den wahrscheinlichkeitstheoretischen Begriffen und Begründungen der Autoren, so kann man letztendlich die Wahrscheinlichkeit, die die Autoren für ihre Sprachmodelle angeben, gleichsetzen mit einer heuristischen Gütefunktion für Modelle.

Betrachtet man die Arbeit im Detail, so ist diese Gütefunktion im wesentlichen uniform, bevorzugt aber Modelle mit wenigen Stämmen und Suffixen. Es ist einsehbar, dass man mit solch einem Ansatz keine vollkommen sinnlose Segmentierung des Textes oder der Wörter bekommen wird, ganz unabhängig davon, was als theoretische Begründung gegeben werden mag: Sucht man in einem natürlichsprachigen Text nach möglichst wenigen sich möglichst häufig wiederholenden Zeichenketten, so kann man erwarten, dass man den Text bis zu einem „gewissen“ Grad in Wörter oder ähnliche linguistisch sinnvolle Einheiten zerlegt hat.[43]

Wahrscheinlich wegen ihrer Bezugnahme auf den Bayesschen Wahrscheinlichkeitsbegriff und vielleicht auch wegen der Tendenz ihres Systems zu eher kompakten Modellen werden Snover und Brent (2001); Snover et al. (2002) oft unter eine sehr wichtige Untergruppe von bayesianischen Arbeiten gerechnet (z.B. von Clark (2001)). Diese sogenannten *Minimum Description Length* (MDL)-Ansätze sind aber doch ein wenig spezifischer zu fassen wie wir in Kürze sehen werden.

Große Beachtung fanden die Forschungen von Creutz und Lagus, die sie seit 2002 in einer Serie von Artikeln publiziert haben **(Creutz, 2003; Creutz und Lagus, 2002, 2004, 2005a,b)** und die sie in **Creutz und Lagus (2007)** in einem geschlossenen Rahmen evaluieren.

Mit Creutz und Lagus kehren wir zur vollen Form des Bayes'schen Gesetzes (Gleichung 2.1) zurück und berechnen die Wahrscheinlichkeit, den Text T aus dem Modell M zu gewinnen, unter der Annahme einer *a priori*-Wahrscheinlichkeitsverteilung der in Betracht gezogenen Modelle. Es ergibt sich eine neue, *a posteriori*-Wahrscheinlichkeitsverteilung über die Modelle. Creutz und Lagus wählen jeweils das Modell mit der größten *a posteriori*-Wahrscheinlichkeit aus. Entsprechend wird dieser

[43] Snover und Brent (2001) sind die einzigen mir bekannten Autoren, die explizit angeben, sowohl Wortlisten und ganze Texte zerlegen zu können („This paper describes a system for unsupervised learning of morphological affixes from texts or word lists.“). Letztlich durchzuführen scheinen sie aber nur die Segmentierung von Wortlisten.

Ansatz als *Maximum A Posteriori* (MAP) bezeichnet.

Parallel untersuchen sie auch eine einfachere Optimierungsstrategie, die dasjenige Modell auswählt, unter dem das tatsächliche Korpus am wahrscheinlichsten ist. In diesem Fall bleibt die *a priori*-Wahrscheinlichkeit der Modelle unberücksichtigt. Dieses als *Maximum Likelihood* bekannte Verfahren zur Modellauswahl war bereits auf Seite 31 Thema.

Das grundlegende Modell, dass Creutz und Lagus in ihren Arbeiten entwickeln, ist von erheblicher Komplexität und beinhaltet insbesondere die Kategorisierung der gefundenen *Morphe*. In diesem Punkt ähnelt die Methode also der von Schone und Jurafsky (2000, 2001) vorgestellten. Diese Kategorisierung implementiert ein *Hidden Markov Modell* unter Verwendung von vier Kategorien: *prefix*, *suffix*, *stem* und *non-morph*.

Diese Kategorien sind allerdings sehr weit und selbst nicht gelernt. Überhaupt ist alles, was Creutz und Lagus (z.B. in Creutz und Lagus (2002, 2007)) an Grammatik in ihre Modelle einbringen, recht heuristisch. Dies ist ein weiteres Beispiel für das Einbringen von sprachlichem Wissen in das Modell selber, ohne, dass die anscheinende Unüberwachtheit direkt durch Positivbeispiele, also einen Goldstandard, verloren ginge.

Creutz und Lagus haben ihr Modell mit Blick auf die agglutinierende Morphologie des Finnischen entwickelt, sie testen es aber auch an Englisch mit konsistenten Ergebnissen.

Für das Hidden Markov Modell verwenden die Autoren die lineare Abfolge der Wörter des Korpus, obwohl sie für die Segmentierung des Korpus diesen bereits als tokenisiert annehmen. Ohne Veränderung sollte der Algorithmus also nicht auf Sprachen anwendbar sein, in denen Tokenisierung nicht trivial ist.[44]

Die verschiedenen Teile des hier nur sehr oberflächlich beschriebenen Systems wurden im Laufe der Zeit von Creutz und Lagus in 4 verschiedenen Ausprägungen implementiert, die sie im Artikel von 2007 vergleichend evaluieren. Das vielleicht erstaunlichste an den berichteten Ergebnissen ist die relative Unempfindlichkeit der Ergebnisse in Bezug auf das im einzelnen verwendete Modell.

Die Autoren vergleichen ihr Modell mit dem soeben bereits erwähnten *Minimum Description Length*-Ansatz zur MI. Goldsmith (2001) ist das Standardbeispiel dieser als MDL abgekürzten Ansätze und hat sich zu einer Art Quasistandard entwickelt. Im folgenden möchte ich auch auf diese Klasse von Verfahren kurz eingehen.

Minimum Description Length (MDL)

Ein kurzes Beispiel soll das hinter diesen Ansätzen stehende Prinzip erläutern: Die Zeichenkette `eine Rose ist eine Rose ist eine Rose` hat 38 Buchstaben (Bytes). Eine kürzere Beschreibung wäre `12312312` zusammen mit dem Wörterbuch `1=eine,2=Rose,3=ist`. In dieser Darstellung braucht der Text nur noch 8 Byte, das Wörterbuch 19. Insgesamt sind es also mit 27 Byte erheblich weniger als für die ausgeschriebene Form des Textes.

Die Beziehung zum bereits ein wenig eingehender diskutierten Begriff der Entropie und dem Konzept der Segmentierung durch Komprimierung ist augenfällig (Vergleiche zum

[44] Womit nicht gesagt werden soll, dass Tokenisierung in den europäischen Alphabetsprachen ohne Probleme ist.

Beispiel Seite 29 ff.). Damit besteht auch wieder eine Beziehung zur Forschungstradition nach Harris.

MDL basiert auf dem Postulat, dass eine Minimierung der kombinierten Länge aus Text und Lexikon die korrekte Liste der *(minimalen) sprachlichen Segmente* eines Textes liefert. Mögliche Begründungen für dieses Postulat werden in den originalen Texten nicht sehr ausgiebig diskutiert. Stattdessen findet sich im Normalfall ein Verweis auf Rissanen (1989). Lediglich einige Autoren wie de Marcken (1996, 39) oder Goldwater et al. (2009, 27) werden konkreter, indem sie MDL explizit als bayesianischen Ansatz mit dem Prior $P(G) = 2^{-|G|}$ beschreiben, wobei das Sprachmodell hier mit G bezeichnet wird (für *grammar*). $|G|$ ist die Größe von G bzw. der Platz, der notwendig ist, um G aufzuschreiben. Diese Form für $P(G)$ ist gleichbedeutend mit einer schnellen Abnahme der Wahrscheinlichkeit größerer bzw. komplexerer Modelle. Im Endeffekt bewirkt die exponentielle Abhängigkeit von der Größe des Sprachmodells eine Bevorzugung möglichst kleiner Modelle.

Dies ist zwar eine Präzisierung der oben skizzierten Idee, begründet diese aber immer noch nicht näher. De Marcken gibt dies zu, indem er schreibt, MDL sei „ just a heuristic“. Goldsmith, der vielleicht bekannteste Vertreter des MDL-Ansatzes schreibt ähnlich: „The idea [...] is based on a simple intuition: that between two extreme analyses, there must be a happy medium that is optimal“. Als extrem bezeichnet er die folgenden beiden trivialen Möglichkeiten, den Text zu kodieren: Auf der einen Seite kann man den Text als Ganzes in das Wörterbuch verlagern. Dies maximiert die Länge des Wörterbuchs, reduziert aber die Länge des Textes auf 1 Byte. Gegenteilig kann man in das Wörterbuch nur die Einzelbuchstaben aufnehmen und den Text in seiner gedruckten Form speichern. Dies ergibt zwar das kleinstmögliche Wörterbuch, maximiert aber die Darstellung des Textes.

Ein Zitat aus Creutz und Lagus (2007, 2) verdeutlicht, welch weiter Bogen geschlagen wird, um eine theoretische Begründung für das MDL-Prinzip zu geben:

> The least-effort principle corresponds to Occam's razor, which says that among equally performing models one should prefer the smallest one. This can be formulated mathematically using the Minimum Description Length (MDL) principle (Rissanen, 1989) or in a probabilistic framework as a maximum a posteriori (MAP) model.

Occam's Razor ist sicherlich ein mächtiges Prinzip, um zwischen konkurrierenden Begriffsgebäuden und Theorien die sparsamste auszuwählen, man sollte aber wohl nicht von vornherein annehmen, dass sich menschliches Sprachvermögen an größtmöglicher Einfachheit orientiert. Das hieße, ein Auswahlverfahren zwischen Modellen mit dem Modell selbst zu verwechseln.

Die ersten Arbeiten, die das MDL-Prinzip anwendeten, um *sprachliche Segmente* zu erkennen, waren wohl Brent (1993); Brent et al. (1995). Aus Platzgründen werde ich auf diese Arbeiten aber nicht näher eingehen. Statt dessen wenden wir uns zwei weiteren wichtigen Beispielen aus der MDL-Familie zu, die beide auch bereits erwähnt wurden: De Marcken (1996) und Goldsmith (2001).

De Marcken (1996) geht über die Mehrzahl der von mir rezipierten Arbeiten hinaus, indem er explizit keine Tokenisierung des Textes voraussetzt, sondern vom Text als

Zeichenkette ausgeht.

Das Verfahren liefert darüber hinaus eine hierarchische Struktur. Beide Punkte sind in der gesichteten MI-Literatur eher eine Seltenheit. Da mein Algorithmus ebenfalls diese Eigenschaften aufweist, werde ich nach dessen Vorstellung noch einmal auf de Marcken's Arbeit zurückkommen (s. Seite 56).

Sein Algorithmus besticht durch Klarheit. Er beginnt mit dem minimalen Lexikon, das genau die Buchstaben als Einträge enthält. Das Verfahren arbeitet iterativ. Ausgehend vom trivialen Anfangszustand werden je zwei Lexikoneinträge zu einem neuen verbunden, wenn dadurch die kombinierte *description length* von Text und Lexikon kleiner wird.

De Marcken gelingt es, einen erheblichen Anteil der nach linguistischen Gesichtspunkten bestimmbaren *sprachlichen Segmente* korrekt zu identifizieren. Allerdings liefert sein Verfahren keinen Anhaltspunkt dafür, welcher Stufe seines Segmentierungsbaums tatsächlich *sprachlichen Segmenten* entsprechen. Die erwähnte hierarchische Struktur beginnt auf Satzebene und endet erst auf *Zeichen*ebene.

Goldsmith (2001) sieht sich selbst in der Nachfolge von de Marcken (1996), ohne dessen Vorarbeit er seine eigenen Forschungen zur MDL nicht gestartet hätte. Auch Harris (1967, 1955) zitiert er prominent und wohlwollend und verwendet seine Ideen explizit in einem ersten Schritt seines Algorithmus. Es zeigen sich also wieder Verbindungen zwischen den bayesianischen Ansätzen und der aus dem Strukturalismus entstandenen Forschungstradition nach Harris.

Ihm geht es sehr viel spezifischer als de Marcken darum, Wörter in Stämme und Suffixe[45] zu zerlegen und darüber hinaus Stämme, die mit denselben Mengen von Suffixen auftreten, zusammenzuordnen. Die entstehenden Klassen nennt er *signatures*, was in diesem Fall ungefähr mit *Paradigmen* zu übersetzen wäre.

Dadurch, dass er ein sehr viel starrer gefasstes Grundmodell für seine Morphologie vorgibt, umgeht er das Problem der starken Übersegmentierung, das bei de Marcken zu beobachten ist.

Dafür tritt er gegenüber de Marcken (1996) von dem Anspruch zurück, Texte als untokenisierte Zeichenketten zu segmentieren. Stattdessen arbeitet er auf der Grundlage von Wortlisten.

Sein Algorithmus beginnt mit einer heuristisch motivierten Anfangsmorphologie. Eine der verwendeten Heuristiken sortiert alle möglichen Suffixe bis zu einer maximalen Länge von 6 nach einer einfachen häufigkeitsbasierten Gütefunktion. Ausgehend von dieser anfänglichen Segmentierung werden – wiederum nach verschiedenen Heuristiken – Veränderungen vorgenommen und akzeptiert, wenn sie die *Description Length* reduzieren. Wie bei vielen Suchalgorithmen gibt es keine Möglichkeit das globale Minimum der *Description Length* mit Sicherheit zu finden. Nur lokale Minima sind auffindbar. Goldsmiths System ist unter dem Namen Linguistica frei verfügbar[46] und wird oft als Referenzsystem verwendet.

Ein Teil der von **Creutz und Lagus** vorgeschlagenen (Creutz, 2003; Creutz und

[45] Wobei ein Stamm wiederum aus Stamm und Suffix bestehen kann.

[46] `http://humanities.uchicago.edu/faculty/goldsmith/Linguistica2000/`

Lagus, 2002, 2004, 2005a,b) MI-Verfahren[47] fällt ebenfalls in die Kategorie der MDL-Ansätze. Creutz und Lagus besinnen sich zurück auf de Marcken (1996) und erlauben eine hierarchische Struktur des Lexikons: ein Lexikoneintrag kann rekursiv aus anderen Lexikoneinträgen[48] bestehen. Dabei begegnen Ihnen ähnliche Probleme wie sie auch im Zusammenhang mit dem von mir vorgestellten Algorithmus zu besprechen sein werden: Die Zusammenordnung von korrekten Segmenten scheitert auf einer höheren Ebene. Sie bringen das Beispiel `[micro[organism s]]` (Creutz und Lagus, 2007, 29).

In der Praxis geben MDL-Verfahren recht gute Ergebnisse. Da ein Text ganz offensichtlich in sich wiederholende Einheiten zerfällt – welchen Status auch immer –, ist dies auch zu erwarten. Weder sind sie aber perfekt, was ja ein gewichtiges Argument wäre, an ihre Wahrheit zu glauben, noch scheitern sie in einem Sinne, dass man ihre Angemessenheit ganz und gar verwerfen könnte. Der Frage nach dem erzeugenden System von Sprache oder dem menschlichen Vermögen, Sprache zu dekodieren, bringen uns solche Untersuchungen also nicht wirklich weiter. Dieser Anspruch schwingt aber durchaus häufig in den entsprechenden Arbeiten mit. So schreibt Goldsmith (2010, 19):

> MDL-based approaches work quite well in practice, and as a selling point, they have the advantage that they offer a principled answer to the question of how and why natural language should be broken up into chunks.

Kritisch urteilt Hammarström (2006, S. 9, Zitate in Format und Sortierung angepasst) über diese Praxis:

> Many publications (Ćavar et al., 2004; Brent et al., 1995; Goldsmith, 2001; Dejean, 1998; Snover et al., 2002; Argamon et al., 2004; Goldsmith et al., 2001; Creutz und Lagus, 2005b; Neuvel und Fulop, 2002; Baroni, 2003; Gaussier, 1999; Sharma et al., 2002; Wicentowski, 2002; González, 2004),
>
> and various other works by the same authors, describe strategies that use frequencies, probabilities, and optimization criteria, often Minimum Description Length (MDL), in various combinations. So far, all these are unsatisfactory on two main accounts; on the theoretical side, they still owe an explanation of why compression or MDL should give birth to segmentations coinciding with morphemes as linguistically defined. On the experimental side, thresholds, supervised/developed parametres and selective input still cloud the success of reported results, which, in any case, aren't wide enough to sustain some too rash language independence claims.

Auch der hier vorgestellte Algorithmus wird keine definitiven Antworten auf theoretische Fragen nach dem grundlegenden Aufbau von Sprache oder der Natur der menschlichen Sprachfähigkeit geben können.

Einerseits aber erhebe ich nicht den Anspruch, meinem Algorithmus einen theoretischen Unterbau zu geben, der sich auf dem jetzigen Stand unseres Wissens doch nicht ohne weiteres aus einer echten Theorie ableiten lassen würde.

[47] zusammengefasst in Creutz und Lagus (2007)

[48] was zu der Terminologie führt, dass ein Morph rekursiv aus weiteren Morphen bestehen kann.

Andererseits aber stellt mein Algorithmus doch den vollständigsten mir bekannten Versuch dar, Frequenzdaten zur Segmentierung von Sprache zu nutzen. Gerade aus den nach wie vor bestehenden Lücken lassen sich Vermutungen ableiten, welche weiteren Zusätze nötig sein könnten, dass der Computer ein wesentliches Stück weiter in die Struktur eines Textes eindringen könnte (s. Abschnitt 2.7).

Hierarchical Bayesian Models

Den fortgeschrittensten Ansatz aus der bayesianischen Familie stellen die sogenannten *Hierarchical Bayesian Language Models* dar (Teh, 2006; Goldwater et al., 2006; Mochihashi und Sumita, 2008; Johnson et al., 2007; Xu et al., 2008; Johnson, 2008; Goldwater et al., 2009; Mochihashi et al., 2009).

In der übrigen Forschung wird diese relativ neue Strömung noch nicht sehr stark wahrgenommen. Erst Goldsmith (2010) bringt eine etwas eingehendere Zusammenfassung. Auch Hammarström (2009) zitiert unter anderem Goldwaters Doktorarbeit (Goldwater, 2007).

Die Sprachmodelle, die in diesen Arbeiten entwickelt werden, bauen auf dem *Chinese Restaurant Process* auf. Dieser stochastische Prozess wird meist mit einer etwas weit hergeholt scheinenden Analogie erklärt: Wir betrachten ein chinesisches Restaurant mit unendlich vielen unendlich großen (runden) Tischen. Nun betreten nach und nach Kunden das Lokal. Der erste setzt sich an einen willkürlich ausgewählten Tisch. Alle folgenden Gäste setzen sich mit einer Wahrscheinlichkeit an bereits besetzte Tische, die zur Zahl der dort bereits sitzenden Gäste proportional ist. Mit einer Restwahrscheinlichkeit werden neue Tische eröffnet.

Dieses Verfahren führt dazu, dass stark besetzte Tische immer mehr Gäste anziehen (*the rich get richer*). Innerhalb des Sprachmodells werden die Tische als linguistische Einheiten interpretiert und die Gäste als deren Vorkommen. Gewonnen hat man damit ein Modell, das in der Lage ist, das Potenzverhalten von Wortfrequenzverteilungen (das Zipf'sche Gesetz) zu reproduzieren.

Das hierarchische des Verfahrens ist, dass die Zusammenfassung von *Zeichen* zu Segmenten nur der erste Schritt ist. Die gefundenen Segmente können im Anschluss wiederum zu Einheiten höherer Ebene zusammengefasst werden.

Die Verwendung eines angemessenen und zudem mathematisch gut durchdachten Modells macht diesen Zweig der MI-Forschung zu ihrem vielleicht vielversprechendsten Ansatz. Auch verwenden diese Autoren nicht heuristisch motivierte ad-hoc Suchverfahren wie viele der bisher vorgestellten bayesianischen Modelle, sondern verlassen sich auf moderne allgemein anerkannte Suchalgorithmen wie den Gibbs-Sampler (Casella und George, 1992) um den riesigen Raum der Modellparameter effektiver zu durchlaufen.

Zusammenfassung

Die vergangenen 60 Jahre haben insgesamt eine große Fülle der vielfältigsten Algorithmen zur *Morphologieinduktion* hervorgebracht. Sie haben gezeigt, dass es bis zu einem gewissen Grad möglich ist, die morphologischen Einheiten, aus denen ein Text beste-

ht, aus oberflächenbasierte Häufigkeiten abzuleiten. Die wichtigsten Strömungen dieser Forschungsrichtung habe ich versucht, zusammenzufassen.

Die grundlegenden Ideen, die hinter den bisherigen Arbeiten zur MI stehen, sind nicht scharf voneinander abgrenzbar und miteinander verwandt. Sie lassen sich ungefähr in folgende Worte fassen: Gute Kandidaten für *sprachliche Segmente* sind Zeichenketten, die häufig zusammen vorkommen und frei kombinierbar sind, oder Zeichenketten an deren Grenzen sich Fortsetzungen schwer vorhersagen lassen.

Es lassen sich zwei grundlegende Strategien unterscheiden, mit deren Hilfe diese Grundideen umgesetzt werden: In vielen Arbeiten werden die häufigkeitsbasierten Prinzipien mittels Heuristiken direkt in Algorithmen implementiert.

In einer anderen Klasse von Algorithmen werden die Prinzipien in Modellfamilien übersetzt. Dabei zählt als gutes Modell eines, das die Daten gut erklärt bzw. eines, das aufgrund der Daten als wahrscheinlich erscheint. Für den Optimierungsprozess bedarf es gewöhnlich eines Suchalgorithmus.

Innerhalb dieser Klasse verdienen die *Hierarchical Bayesian Models* besondere Beachtung. Sie sind mathematisch fundiert formuliert und die Struktur der Modelle steht mit den bekannten statistischen Eigenschaften von Texten in guter Übereinstimmung.[49]

Die folgenden grundlegenden Probleme tauchen in vielen Arbeiten auf:

Die gegebenen theoretischen Begründungen sind schwach. Es gibt keinen wirklichen theoretischen Überbau, aus denen sie folgen würden. Die realweltliche Verankerung der eingehenden Begriffe wie *Entropie* und *Wahrscheinlichkeit* im System der Sprache ist schwach oder wird nicht thematisiert. Die Daten, die sich ergeben, sind nicht geeignet, über die theoretischen Voraussetzungen zu urteilen.

Falls Sprachmodelle eingesetzt werden, bedarf es immer eines mehr oder weniger komplexen Suchalgorithmus. Die genaue Performanz dieser Suche, ihr Erfolg beim Auffinden globaler Optima und die Struktur des Suchraumes werden oft nicht untersucht.

Die einfache Tatsache, dass längere Zeichenketten notwendigerweise seltener sind führt für die verwendeten Maße in der Regel ebenfalls zu einem tendenziellen Abfall. Dieser Abfall macht die Verwendung von vielen Kriterien zur Segmentierung fehleranfällig. Das Problem wird nicht oft benannt. Eine quantitative Bearbeitung ist mir unbekannt.

2.5 Der Algorithmus

Im folgenden Kapitel stelle ich selbst einen Algorithmus vor, der die in Abschnitt 2.3 definierte Aufgabe zu lösen sucht. Die grundlegende Idee dieses Algorithmus ist die Präzisierung eines in der *Morphologieinduktion* allgegenwärtigen Prinzips:

[49] Ich betone an dieser Stelle noch einmal, dass das Zitieren und Vergleichen von Performanzwerten keinen Mehrwert bringt, da ihre jeweiligen Grundlagen zu unterschiedlich sind. So berichten Goldwater et al. (2009) durchaus sorgfältig erhobene Werte. Sie stellen auch über die Wahl von Korpus und Evaluationsmethode Vergleichbarkeit mit Brent (1999); Venkataraman (2001) her. Allerdings verhindern sowohl die spezielle Wahl des Korpus (die CHILDES Datenbank (MacWhinney und Snow, 1985) mit an Kleinkinder gereichtete Äußerungen) als auch das genaue Ziel der Analyse (Wiederfinden der Wortgrenzen in den Äußerungen) eine Vergleichbarkeit mit jeweils allen anderen Arbeiten.

> *Segment*kandidaten sind Zeichenketten, deren Endpunkte besser vorhersagbar sind als ihre Verlängerungen.

Im Gegensatz zu den mir bekannten veröffentlichten Algorithmen besteht die Datengrundlage aus allen Substringhäufigkeiten im Trainingstext.

Ich beginne mit einem Beispiel, dass sich durch die gesamte Beschreibung des Algorithmus ziehen wird. Sei das Fragment

$t =$`he_has_accomplished_some`

der Testtext. Ihn gilt es zu segmentieren. Als Trainingstext T verwenden wir das Brownkorpus (Francis und Kucera, 1967).[50] Die Häufigkeiten aller Zeichenketten dieses Testtextes ist in Abbildung 2.1 dargestellt.

Man erkennt zusammenhängende Gebiete größerer Häufigkeit. So ist das Wort `accomplished` anhand des dunkler grauen Gebiets zu erkennen, dass in seiner rechten oberen Ecke durch einen Punkt gekennzeichnet ist. Man meint bereits in diesem kleinen Datenausschnitt viele solcher Strukturen zu erkennen.

Der erste Schritt bei der Entwicklung des Algorithmus wird es sein, das eingangs formulierte Prinzip in Bezug auf die dargestellten Häufigkeitsdaten zu formalisieren.

Ebenfalls kann man aber in Abbildung 2.1 erkennen, dass sich die dunkleren Bereiche größerer Häufigkeit massiv überlagern. Daher wird jede Formulierung aufgrund lokaler Häufigkeitsdaten notwenig zu einer hohen Mehrdeutigkeit führen. Es bedarf also einer Disambiguierungsstrategie.

Die Disambiguierung teilt sich in zwei Schritte: Die große Masse der irreführenden Segmentkandidaten[51] wird durch die Forderung eliminiert, dass der Satz aus einer lückenlosen Aneinanderreihung von Segmenten bestehen muss.

Die verbleibenden Mehrdeutigkeiten werden durch größtenteils häufigkeitsbasierte Heuristiken aufgelöst, die für konkurrierende Möglichkeiten eine Gütereihenfolge festlegen. Dies führt zu einer scheinbaren Vielfalt an Algorithmusvarianten, aus denen erst eine einzige ausgewählt werden muss, um wieder zu einem unüberwachten Algorithmus zu gelangen.

Im empirischen Teil (2.6) wird sich zeigen, dass es genau diese Vielfalt ist, aus der sich bei genauer Betrachtung und sorgfältiger Auswertung die linguistisch interessanten Schlüsse ziehen oder auf deren Grundlage sich weiterführende Fragen formulieren lassen. Diese Reichhaltigkeit der Daten eröffnet im günstigsten Fall einen empirischen Zugang zu theoretisch relevanten Fragestellungen zum erzeugenden System der Texte (Sprache), der in den bisher veröffentlichen Arbeiten meist fehlt.

Ich gebe für den vorgestellten Algorithmus keine theoretische Begründung. Damit bleibt die gesamte Untersuchung explorativ. Genau dies sehe ich als Vorteil, da fragwürdige theoretische Fundierungen den Blick auf die Daten eher verstellen als Einsichten in ihre Struktur zu ermöglichen. Und fragwürdig muss jede theoretische Fundierung

[50] Bzw. eine um die letzten 500 Sätze (0.1%) gekürzte Version. Der fehlende Teil wurde als Testmaterial zurückbehalten.

[51] Dies könnte `lished_` sein.

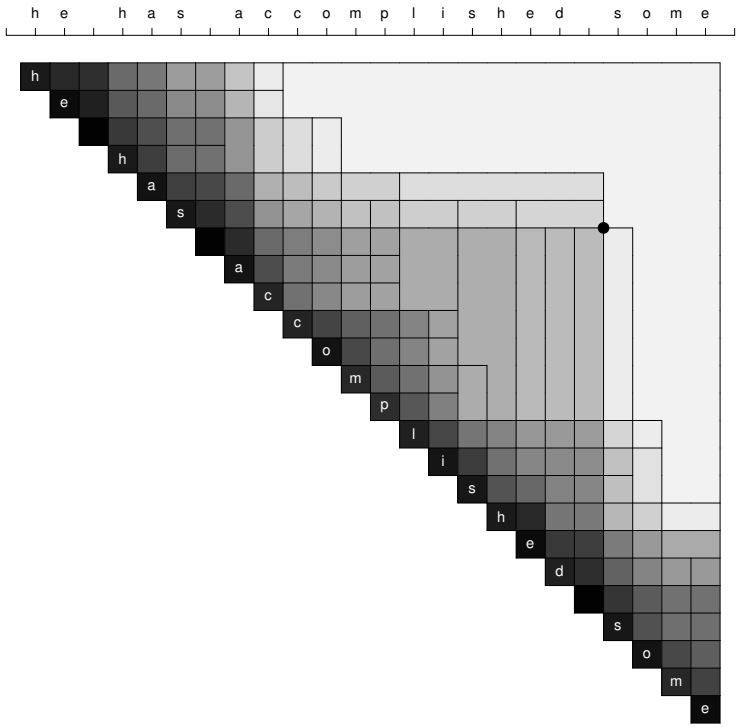


Abbildung 2.1: Bildliche Darstellung aller Häufigkeiten der Substrings des Beispieltextes `he has accomplished some`. Dieser ist zur Verdeutlichung nicht nur auf der x-Achse, sondern noch einmal auf der Diagonalen eingetragen. Als Referenz- bzw. Trainingstext diente eine leicht verkürzte Version des Brownkorpus (Francis und Kucera, 1967). Die Häufigkeiten der einzelnen Strings sind als Helligkeiten kodiert: Je dunkler ein Feld, desto häufiger ist der entsprechende String. Diesen kann man ablesen, indem man von einem Punkt der Graphik waagerecht nach links und senkrecht nach unten geht. So entspricht der Punkt in der rechten oberen Ecke des grauen Bereiches der Häufigkeit des gesamten Textes, bzw. Satzfragmentes, nämlich 1. Das Feld, aus dem sich die Häufigkeit der Zeichenkette `_accomplished_` ergibt, ist durch einen schwarzen Punkt gekennzeichnet. Zur Verdeutlichung trennen schwarze Begrenzungen Felder (Strings) mit verschiedenen Häufigkeiten. Geht man von der rechten oberen Ecke nach links, so ändert sich erst einmal nichts an der Häufigkeit der entsprechenden Strings: Auch der verkürzte Text `he has accomplished som` kommt im Browncorpus nur einmal vor. Erst `he has ac` kommt 2 mal vor. Die Felder auf der Diagonalen entsprechen folgerichtig der Häufigkeit der einzelnen *Zeichen*: Das Leerzeichen ist am häufigsten (966311), gefolgt vom `e` (584742).

sein, solange es keine etablierte Theorie für das spracherzeugende System gibt, das als ausreichend empirisch fundiert gelten kann.

Ein Vorteil des Algorithmus liegt auch im Fehlen eines Suchalgorithmus. Dies vereinfacht einerseits die Berechnungen erheblich und sichert andererseits die Eindeutigkeit und Vollständigkeit der Ergebnisse: Die gewählte Algorithmusvariante führt bei gleichen Daten eindeutig zu identischen Ergebnissen, ohne Rücksicht auf eine Optimierungsstrategie.

Das Abfallproblem taucht im vorgestellten Algorithmus nicht mehr auf. Diese Eigenschaft im Vergleich zu einigen anderen in der Literatur vorgeschlagenen Algorithmen wird empirisch zu belegen sein, wenn das Verfahren ausreichend genau beschrieben wurde (Abbildung 2.8 und Diskussion dazu).

2.5.1 Die Identifizierung konkurrierender Segmentierungen

Der soeben dargestellte Ausgangspunkt des Algorithmus ähnelt der Grundidee, auf der schon Harris (1955) seinen Algorithmus aufgebaut hat, vergleiche die Diskussion auf Seite 28 f. und Abschnitt 2.4.1: Ich beginne mit der Beobachtung, dass sich der letzte Buchstabe eines Wortes oder genauer eines *sprachlichen Segmentes* (Definition 6) oft sehr leicht vorhersagen lässt, während dies für das unmittelbar auf das Wort oder Segment folgende *Zeichen* nicht gilt: In einem englischen Text wird nach der Zeichenkette `biolo` im Normalfall ein `g` folgen (Brownkorpus: in allen 33 Fällen). Danach spalten sich die Möglichkeiten auf in ein `y` (7 Fälle) für `biology` und ein `i` (26 Fälle) für `biolog{ic,ically,ist,ists,ical}`. Denselben Abfall der Vorhersagbarkeit kann man in der anderen Richtung für den Wort/Segmentanfang beobachten.

Um diese Grundidee zu einem umsetzbaren Algorithmus formen zu können, ist es in einem ersten Schritt notwendig, die informelle Beschreibung der „Vorhersagbarkeit“ und ihres Abfalls zu quantifizieren und zu formalisieren.

An dieser Stelle sei noch einmal daran erinnert (s. Bemerkung auf Seite 16 nach Definition 3), dass es im vorliegenden Kontext keinen prinzipiellen Unterschied zwischen Text und Korpus gibt. Korpora im linguistischen Sinne werden dem Algorithmus zu Texten verknüpft übergeben. Praktisch ist es so, dass der Trainingstext oder das Trainingskorpus als ein einziger längerer Text vorliegt, während der angebotene Testtext jeweils aus einem Satz oder Absatz des gesamten Testmaterials besteht.

Ich kehre zurück zum in Abbildung 2.1 visualisierten Beispiel. Betrachten wir zunächst den Substring `_accomplish`. Seine Frequenz ist 87. Die um einen Buchstaben kürzere Zeichenkette `_accomplis` kommt im Trainingstext ebenfalls 87 mal vor. Das `e` ist also gut *vorhersagbar*. Formal definiere ich diesen Begriff wie folgt:

Definition 14 (forward predictability) *Sei $s_{ij}^t = t_i t_{i+1}..t_j$ mit $1 \leq i \leq j \leq n$ eine Zeichenkette des Testtextes $t = t_1 t_2..t_n$. Das Trainingskorpus sei mit T bezeichnet. $N_T(s_{ij}^t)$ sei*

- *für $i \leq j$ die Zahl der Vorkommen von s_{ij}^t in T, zuzüglich eins.*[52]

[52] Das Addieren von 1 ist eine Designentscheidung bei der Entwicklung des Algorithmus gewesen und

- *für $i > j$ die Gesamtzahl der Zeichen in T, auch bezeichnet mit $L(T)$.*

Dann ist für $i > 1$ und $j < n$ die forward predictability des Zeichens t_{j+1} nach dem String s_{ij}^t relativ zum Trainingstext T

$$V_T^+(s_{ij}^t, t_{j+1}) = \frac{N_T(s_{ij+1}^t)}{N_T(s_{ij}^t)} \tag{2.2}$$

Diese auf den ersten Blick komplizierte Definition übersetzt sich für unser Beispiel in einen sehr einfachen Ausdruck. Die *forward predictability* `h` nach `_accomplis` ist

$$V_T^+(s_{7\,16}^t, t_{17}) = V_T^+(\texttt{\_accomplis}, \texttt{h}) = \frac{N_T(\texttt{\_accomplish})}{N_T(\texttt{\_accomplis})} = \frac{87}{87} = 1 \ ,$$

Die Fallunterscheidung für $N(s_{ij})$ für $i \leq j$ und $i > j$ in Definition 14 macht diese und die folgenden Definitionen auch anwendbar auf Unigramme. So ist z.B. die *forward predictability* eines *Zeichens* ohne Kontext gegeben als

$$V_T^+(s_{i,i-1}^t, t_i) = \frac{N_T(t_i)}{N_T(s_{i,i-1}^t)} = \frac{N_T(t_i)}{L(T)} \tag{2.3}$$

So ergibt sich für das `h` in `accomplished` ohne Kontext die Vorhersagbarkeit

$$V_T^+(s_{17\,16}^t, t_{17}) = \frac{N_T(t_{17})}{N_T(s_{17\,16}^t)} = \frac{N_T(\texttt{h})}{L_T} = \frac{247153}{5948881} = 0.042$$

Derselbe Wert ergibt sich natürlich auch für jedes andere `h` im Testtext.

Es sei hier betont, dass die *Vorhersagbarkeit* quantifiziert, in welchem Ausmaß der nächste tatsächlich im *Testtext* vorkommende Buchstabe vorhergesagt werden kann. Entropiebasierte Ansätze wie die erwähnten Arbeiten von Cohen et al. (2007); Hafer und Weiss (1974) neigen statt dessen dazu, die Verteilung aller möglichen, dh. aller im *Trainingskorpus* vorkommenden Fortsetzungen zu betrachten. Dieser meines Erachtens konzeptuell irreführende Weg korrespondiert auffällig mit der Neigung dieser Ausrichtung der relevanten Forschung, relative Häufigkeiten mit Wahrscheinlichkeiten gleichzusetzen. „Vorhersagbarkeit“ dagegen ist nur eine Metapher zur Beschreibung der Häufigkeitsverhältnisse im Trainingstext.

Die *backward predictability* von t_{i-1} vor s_{ij}^t ist analog zur Definition 14 der *forward predictability* definiert als

Definition 15 (backward-predictability)

$$V_T^-(t_{i-1}, s_{ij}^t) = \frac{N_T(s_{i-1j}^t)}{N_T(s_{ij}^t)} \tag{2.4}$$

hat keinen Einfluss auf die Ergebnisse.

Für unseren Testtext heißt das zum Beispiel

$$V_T^+(t_7, s_{8\,17}^t) = V_T^+(\texttt{\_}, \texttt{accomplish}) = \frac{N_T(\texttt{\_accomplish})}{N_T(\texttt{accomplish})} = \frac{87}{87} = 1 \ ,$$

dh., in diesem Fall ist die Vorhersagbarkeit in beiden Richtungen maximal.

In der Einführung dieses Abschnittes wurde schon motiviert, dass es nicht direkt die Vorhersagbarkeit des nächsten Buchstabens alleine ist, die es ermöglicht, den Text an sinnvollen Stellen zu segmentieren, sondern die Änderung der Vorhersagbarkeit von Buchstabe zu Buchstabe.

Diesen Abfall sehen wir auch in unserem Beispiel: während der String `_accomplish` 87 mal im Trainingskorpus vorkommt, erscheint das verlängerte $s_t(7, 18) =$`_accomplishe` nur 44 mal.

Dies entspricht einer *forward predictability* von nur noch

$$V_T^+(\texttt{\_accomplish}, \texttt{e}) = \frac{44}{87} = 0.51$$

Diese Änderung der Vorhersagbarkeit gilt es zu formalisieren. Der *forward predictability change* für t_{j+1} nach s_{ij}^t wird definiert als:

Definition 16 (forward predictability change) *Der forward predictability change D_T^+ eines Zeichens t_{j+1} nach einer Zeichenkette s_{ij} relativ zu einem Trainingstext T ist*

$$D_T^+(s_{ij}^t, t_{j+1}) = \frac{V_{T,t}^+(s_{ij}^t, t_{j+1})}{V_{T,t}^+(s_{ij-1}^t, t_j)} = \frac{\frac{N_T(s_{ij+1}^t)}{N_T(s_{ij}^t)}}{\frac{N_T(s_{ij}^t)}{N_T(s_{ij-1}^t)}} = \frac{N_T(s_{ij-1}^t) N_T(s_{ij+1}^t)}{N_T^2(s_{ij}^t)} \tag{2.5}$$

Der *forward predictability change* von `e` nach `_accomplish` ist somit

$$\begin{aligned} D_T^+(s_{7\,16}^t, t_{17}) &= D_T^+(\texttt{\_accomplish}, \texttt{e}) \\ &= \frac{N_T(\texttt{\_accomplis}) N_T(\texttt{\_accomplishe})}{N_T(\texttt{\_accomplish})} \\ &= \frac{87 \cdot 44}{87^2} = \frac{0.51}{1} = 0.51 \end{aligned}$$

Intuitiv bedeutet ein *foward predictability change* kleiner als 1, dass das folgende Zeichen nicht so leicht vorhersagbar ist wie das zuletzt gesehene.

Der *backward predictability change* für t_{i-1} vor s_{ij}^t ist entsprechend:

Definition 17 (backward predictability change)

$$D_T^-(t_{i-1}, s_{ij}^t) = \frac{V_T^-(t_{i-1}, s_{ij}^t)}{V_T^-(t_i, s_{i+1j}^t)} = \frac{\frac{N_T(s_{i-1j}^t)}{N_T(s_{ij}^t)}}{\frac{N_T(s_{ij}^t)}{N_T(s_{i+1j}^t)}} = \frac{N_T(s_{i-1j}^t) N_T(s_{i+1j}^t)}{N_T^2(s_{ij}^t)} \tag{2.6}$$

In vielen Zusammenhängen ist es im Folgenden nicht erforderlich, alle Indizes mitzuziehen. Vereinfacht schreibe ich auch Abkürzungen wie $D^+(s)$ oder gar D^+ anstelle von $D^+(s^t_{ij}, t_{j+1})$ ohne weiter darauf hinzuweisen.

In der MI-Community ist es relativ selten, dass mit der Änderung von Vorhersagbarkeiten, also je nach Sichtweise von relativen Häufigkeiten oder von Wahrscheinlichkeiten gearbeitet wird. So verwenden Hafer und Weiss (1974) beispielsweise sogar direkt absolute Häufigkeiten anstelle von relativen Häufigkeiten oder deren Änderung. Die überwiegende Anzahl der Forscher folgt ähnlichen Mustern.

Nun ist nicht nur der *forward predictability change* von `e` nach `_accomplish` kleiner als eins, sondern auch der *backward predictability change* von `s` vor `_accomplish`:

$$D^-_T(t_6, s^t_{7\,17}) = D^-_T(\texttt{s}, \texttt{\_accomplish}) = \frac{13 \cdot 87}{87^2} = 0.15$$

Wir beobachten also, dass zumindest in unserem Beispiel ein *sprachliches Segment* an beiden Seiten von einem Abfall der Vorhersagbarkeit begrenzt wird. Diese Beobachtung dient uns umgekehrt dazu, in einem weiteren Schritt *mögliche Segmente* zu identifizieren:

Definition 18 (mögliches Segment) *Die Zeichenkette s^t_{ij} aus t ist ein mögliches Segment relativ zu T, wenn und nur wenn*

1. *$i = 1$ oder $D^-_T(t_{i-1}, s^t_{ij}) < 1$ und*
2. *$j = n$ oder $D^+_T(s^t_{ij} t_{j+1}) < 1$.*

i und j heißen Grenzen des möglichen Segmentes s^t_{ij}. Genauer ist i der Anfangspunkt von s_{ij}, j ist sein Endpunkt.

Nach dieser Definition ist $s^t_{7\,17}$ =`_accomplish` ein *mögliches Segment*. Abbildung 2.2 zeigt die aus Abbildung 2.1 bekannten Daten zusammen mit allen *möglichen Segmenten*.

Ando und Lee (2003)[53] verwenden übrigens tatsächlich ein ähnliches Maß wie das von mir vorgeschlagene, nämlich den Abfall von relativen Häufigkeiten. Leider vollziehen sie aber nicht den Übergang von der **Segmentgrenze** zum vollständigen **Segment**.

Der Grund für die vorsichtige Terminologie *mögliches Segment* wird in Kürze klar.

Das auf `_accomplishe` folgende `d` ist wegen N_T(`_accomplished`) = 43 wieder besser vorhersagbar als das `e`:

$$D^+_T(s^t_{7\,17}, t_{18}) = D^+_T(\texttt{\_accomplishe}, \texttt{d}) = \frac{43 \cdot 87}{44^2} = 1.93$$

Dasselbe gilt wiederum auch in der anderen Richtung:

$$D^-_T(t_5, s^t_{6\,17}) = D^-_T(\texttt{a}, \texttt{s\_accomplish}) = \frac{6 \cdot 87}{13^2} = 3.1$$

[53] vgl. auch die Diskussion dieser Arbeit auf Seite 36.

Daher ist weder `_accomplishe` noch `s_accomplish` ein *mögliches Segment*. Nach `_accomplished_` allerdings fällt die Vorhersagbarkeit erneut ab und wir bekommen ein zweites *mögliches Segment* $s_{7\,20}$ = `_accomplished_`:

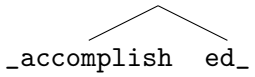
$$D_T^+(\texttt{\_accomplished\_}, \texttt{s}) = \frac{2 \cdot 43}{41^2} = 0.051$$

Um nichts zu vergessen versichern wir uns noch, dass auch für das so verlängerte Segment gilt $D^-(\texttt{s}, \texttt{\_accomplished\_}) = \frac{9 \cdot 41}{41^2} = 0.22 < 1$.

Das heißt, sowohl `_accomplish`, also auch `_accomplished_` wurden bisher als *mögliche Segmente* identifiziert. Was ist mit der Endung `ed_`? Auch hier stellen wir fest, dass der *Predictability change* in beiden Richtungen kleiner ist als 1:

$$D_T^+(\texttt{ed\_}, \texttt{s}) = 0.45 \quad \text{und}$$
$$D_T^-(\texttt{h}, \texttt{ed\_}) = 0.11$$

Insgesamt ergibt sich also die Aufteilung



als mögliche Zerlegung, was ja auch der linguistischen Theorie entspricht. In Abbildung 2.2 ist diese Analyse mit drei größeren Kreisen gekennzeichnet.

Es ist demnach auf dem dargestellten Weg nicht nur möglich, einen Text in Einheiten zu zerlegen, sondern er ermöglicht auch eine teilweise Einordnung dieser Einheiten in hierarchische Strukturen. Diese erstrebenswerte Eigenschaft wurde bereits in Abschnitt 2.3 angekündigt, in dem die gestellte Aufgabe näher charakterisiert wurde.

Wenn man sich das dargestellte Prinzip durchdenkt, kommt man zu dem Schluss, dass unter den kleinsten mit dieser Methode auffindbaren Einheiten Di- oder Trigraphe sein sollten, die gemeinsam ein Phonem darstellen, aber kein *sprachliches Segment* sind (vgl. die Bemerkung vor 4). Ein Beispiel ist das deutsche `sch`, das das Phonem [ʃ] repräsentiert. Das folgende Beispiel bezieht sich auf die groß geschriebene Variante `_Sch` einschließlich des führenden Leerzeichens. Im Deutschen erscheint nach beinahe jedem `_Sc` ein `h` (in Bebel (2004a) in 99.93% der Fälle[54]) wonach die Vorhersagbarkeit stark abfällt (41% für das häufigste `_Schw`). In anderen Worten: $D^+(\texttt{\_Sch}, x) < 1$ für alle möglichen Fortsetzungen x. Dies sieht auf den ersten Blick so aus, als wäre es ein Gegenargument gegen das vorgestellte Segmentierungsverfahren, das es sich ja zum Ziel gesetzt hat, Texte in *sprachliche Segmente* zu zerlegen. `_Sch` ist aber keines, unabhängig davon, ob man das Leerzeichen mitzählt, oder nicht.

Das folgende Beispiel zeigt, dass das Problem in vielen Fällen gar nicht erst auftreten wird: Sei

```
Die_Schule_ist_aus.
```

[54]Dies sind alle bis auf einen. Diese Ausnahme ist eine Erwähnung des schottischen Autors Walter Scott.

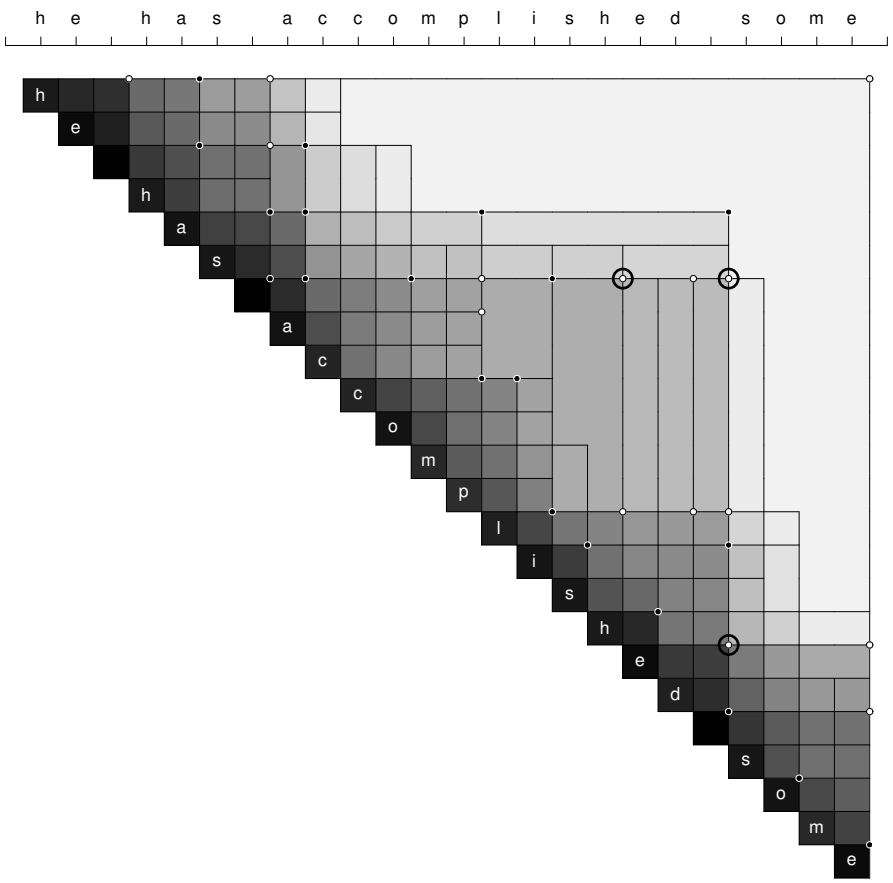


Abbildung 2.2: Dieselben Daten in derselben Darstellung wie in Abbildung 2.1. *Mögliche Segmente* sind durch kleine Punkte in den Feldecken gekennzeichnet. Die Punkte mit weißem Zentrum entsprechen dabei *Segmenten* nach Definition 20, während die schwarz gefüllten Punkte zwar *mögliche Segmente* nach Definition 18 sind, aber keine *Segmente*. Die auf Seite 53 hergeleitete Analyse von `_accomplished_` ist mit drei größeren Kreisen gekennzeichnet. Es ergibt sich auch, dass einzelne *Zeichen mögliche Segmente* sein können. So erhält man für das zweite Leerzeichen: $D^{+}(\texttt{\_}, \texttt{a}) = \frac{L(T)N_T(\texttt{\_a})}{N(\texttt{\_})^2} = \frac{5948881 \cdot 105366}{966311} = 0.67$ und nach ähnlicher Rechnung $D^{-}(\texttt{s}, \texttt{\_}) = 0.69$. $L(T)$ ist wiederum die Länge des Trainingstextes. Es ist übrigens nicht so, dass alle *möglichen Segmente* vom Algorithmus überhaupt betrachtet werden müssen. Von den 21 *möglichen Segmenten*, die keine *Segmente* sind, werden 12 niemals in Betracht gezogen.

der zu zerlegende Testtext. `_Sch` ist ein mögliches Segment, da auch $D_T^-(\texttt{e}, \texttt{\_Sch}) < 1$. Allerdings ist `ule_` kein mögliches Segment im Sinne der oben dargestellten Definition, da $N_T(\texttt{hule\_}) = 18$, $N_T(\texttt{ule\_}) = 19$ und $N_T(\texttt{le\_}) = 1135$, womit sich insgesamt ein *backward predictability change* von $D_T^-(\texttt{h}, \texttt{ule\_}) = \frac{18/19}{19/1135} = 56$ ergibt, dh., obwohl das `ule_` nach dem `_Sch` nicht gut vorhersagbar ist, kann umgekehrt aus dem Vorkommen von `ule_` fast sicher auf das `h` (und auf `_Schule_` insgesamt) geschlossen werden. Auch diese Verhältnisse spiegeln die linguistische Wirklichkeit angemessen wieder: `ule_` ist – zumindest in diesem Zusammenhang – sicher kein *sprachliches Segment*.

Nach der Identifizierung *möglicher Segmente* folgt nun der Schritt zur tatsächlichen *Segmentierung* ganzer Texte bzw. Sätze. Die eben dargestellte Diskussion macht deutlich, dass von einer Segmentierung des gesamten Textes nicht nur zu verlangen ist, dass sie aus *möglichen Segmenten* gemäß der obigen Definition besteht. Darüber hinaus muss sie eine vollständige Zerlegung des Textes in solche *möglichen Segmente* sein.

Bevor ich den Begriff der *Segmentierung* definiere, gilt es noch ein weiteres Detail zu berücksichtigen. Im obigen – englischen – Beispiel gibt es einen *forward predictability change* nach `_has_`, als auch einen *backward predictability change* vor `_accomplish`. Das von beiden Zeichenketten geteilte Leerzeichen dagegen ist in beide Richtungen gut vorherzusagen. Aus diesem Grund wird zugelassen, dass Leerzeichen zu zwei aufeinanderfolgenden Segmenten gehören können. Diese Sonderrolle des Leerzeichen ist eine der wenigen in 2.3 angesprochenen Stellen, an der so etwas wie linguistisches Wissen in den Algorithmus einfließt.

Zusammengefasst ist eine gültige Zerlegung durch folgende Eigenschaften charakterisiert:

Definition 19 (Segmentierung) *Eine Segmentierung S eines Testtextes $t = t_1 \dots t_n$ relativ zu einem Trainingskorpus T ist eine Menge aus k möglichen Segmenten s^t relativ zu T mit folgenden Eigenschaften:*

1. *Die Zerlegung ist vollständig, dh. für jedes $s = s_{ij}^t \in S$ gilt:*
   a) *Entweder ist $i = 1$ oder es existiert ein $s_{vw}^t \in S$ mit $w = i$ oder mit $w = i + 1$, wobei dann t_i das Leerzeichen ist.*
   b) *Entweder ist $j = N$ oder es existiert ein $s_{op}^t \in S$ mit $o = j$ oder mit $o = j - 1$, wobei dann t_j das Leerzeichen ist.*
2. *Falls s mehrere andere mögliche Segmente überspannt, so muss sein Anfangspunkt i mit seinem Endpunkt j durch genau 2 andere mögliche Segmente $s_y = s_{im}^t$ und $s_z = s_{m'j}^t$ verbunden sein, wobei gilt: $m' = m$ oder $m' = m - 1$ und t_m ist das Leerzeichen. s_y und s_z heißen die Kinder von s_x.*

In Abbildung 2.3a sind alle mit dieser Definition kompatiblen Zerlegungen visualisiert.

Das hier vorgestellte Konzept der *Segmentierung* ist im Grunde die Ausformulierung und Formalisierung einer Idee, wie sie zum Beispiel auch bei Bauer (2003, 11) zu finden ist: Ein Text kann in Wortformen zerlegt werden, indem man ihn lückenlos in wiederkehrende Zeichenketten zerlegt. Es ist ein wichtiger Aspekt dieser Arbeit

nachzuprüfen, bis zu welchem Grad dieses einfache Prinzip wirklich ausreichen kann, Texte in Wörter oder allgemein in linguistische Einheiten zu zerlegen.

Die in Abschnitt 2.4.2 besprochene Arbeit von de Marcken (1996) ist in der Lage, eine in der Struktur recht ähnliche Segmentierung zu produzieren wie die hier vorgeschlagenen *Segmentierungen*. Die dabei allerdings auftretende Übersegmentierung wurde bereits erwähnt. Ihre Ursache liegt darin, dass de Marcken keinerlei Bedingungen formuliert, was für Eigenschaften ein Segment haben muss.

Im hier vorgestellten Algorithmus wird der Gefahr der Übersegmentierung durch die Restriktionen auf *mögliche Segmente* mit ihrem beidseitigen Vorhersagbarkeitsabfall und der Lückenlosigkeit von *Segmentierungen* weitgehend reduziert.

Sehr viele Ansätze zum unüberwachten Lernen von Morphologie arbeiten ohne die in einem Text vorkommende Kontextinformation explizit zu verarbeiten. Creutz und Lagus (2007, 4) zitieren[55] in diesem Zusammenhang zum Beispiel Ando und Lee (2003); Yu (2000); Peng und Schuurmans (2001). Diese Einschränkung gilt nicht für den hier vorgestellten Algorithmus. Dies aus zwei Gründen: Zum einen werden n-Gramme unbeschränkter Länge verwendet. An manchen Stellen im Text wiederholen sich sehr lange Zeichenketten. All diese Information kann zur Segmentierung beitragen. Zum anderen werden von vorneherein nur Segmentierungen zugelassen, die den ganzen Satz zu segmentieren erlauben. Wir werden im nächsten Kapitel sehen wie die in allen einen Satz betreffenden Häufigkeiten steckende Information zur Disambiguierung konkurrierender Segmentierungen herangezogen wird.

Die in Definition 19 vorgenommene Festlegung auf genau zwei mögliche Kinder eines *möglichen Segmentes* kann als Einführung weiteren sprachlichen Wissens missverstanden werden. Im gegenwärtigen Algorithmus stellt sie schlicht eine rechentechnische Beschränkung dar. Einer Verallgemeinerung auf mehr als zwei Kinder steht kein prinzipielles Hindernis entgegen. Die Interpretation als sprachliches Wissen wäre natürlich auch deswegen irreführend, weil es auf allen Ebenen linguistischer Analyse Bäume mit mehr als zwei Kindelementen gibt (in vielen Theorien zumindest).

Es kann hilfreich sein, explizit eine weitere Definition aufzustellen:

Definition 20 (Segment) *Mögliche Segmente, die Teil einer Segmentierung sind, heißen Segmente.*

Diese Terminologie unterscheidet die vom Algorithmus vorgeschlagenen *Segmente* von den in Definition 6 eingeführten *sprachlichen Segmenten*, der linguistischen Wirklichkeit (oder Theorie, je nach Sichtweise).

Ich sage, dass ein *Segment* falsch ist, wenn es kein *sprachliches Segment* wie nach Definition 6 ist. Dies trifft im Beispiel wohl unter anderem auf das in Abbildung 2.3 zu erkennende Segment `accomp` zu. *Segmente*, die nur aus dem Leerzeichen bestehen, sollen nicht als *falsch* gelten.

Da nicht jedes *mögliche Segment* Teil einer *Segmentierung* sein kann, ergibt sich eine klarere Terminologie, wenn ihr Kandidatenstatus von Beginn an durch die Zusatzbezeichnung *möglich* explizit gemacht wird.

[55] Statt Ando und Lee (2003) zitieren sie allerdings ein gleichnamiges Werk von 2000.

Man kann sich eine *Segmentierung* vorstellen als eine lineare Abfolge von *Segmenten*, von denen jedes einen Baum bilden kann, in dem jedes Knotensegment zwei Kinder hat oder keines. *Segmente* ohne Kinder nennen wir *Blätter* oder *leaves*. Ich spreche auch von *Segmentbäumen*, wenn der Baumcharakter der hierarchischen Strukturen aus Eltern- und Kindsegmenten betont werden soll.

Ein paar weitere Definitionen bzw. Formalisierungen werden für die folgende Diskussion hilfreich sein:

Definition 21 (kids(s)) *Sei s ein Segment im Sinne von Definition 20. Die Funktion $kids(s)$ bildet s auf die Menge seiner Kindsegmente ab:*

$$kids(s) = \begin{cases} \{s_x, s_z\}, & \text{wenn } s_{x,y} \text{ wie in Definition 19 (Punkt 2) existieren.} \\ \emptyset & \text{sonst} \end{cases}$$

Kurz gesagt, $kids(s)$ sind die Kinder von s. In der in Abbildung 2.3 dargestellten Segmentierung gilt $kids(\texttt{\_accomplished}) = \{\texttt{\_accomplish}, \texttt{ed\_}\}$.

Definition 22 (leaves(s)) *Sei s ein Segment im Sinne von Definition 19. Die Funktion $leaves(s)$ bildet s ab auf die Menge*

$$leaves(s) = \begin{cases} \{s\}, & \text{wenn } kids(s) = \emptyset \\ \{leaves(s_x), leaves(s_y)\} & \text{sonst, wobei } \{s_x, s_y\} = kids(s) \end{cases}$$

Kurz gesagt, $leaves(s)$ gibt die *Blätter* unterhalb eines *Segmentes* s zurück. Ihre Zahl wird mit $|leaves(s)|$ bezeichnet. Im Beispiel ist $leaves(\texttt{\_accomplished\_}) = \{\texttt{\_accomp}, \texttt{lish}, \texttt{ed\_}\}$ und $|leaves(\texttt{\_accomplished\_})| = 3$.

Definition 23 (successor) *Ein Segment $s_{op} \in S$ heißt successor eines anderen Segmentes $s_{ij} \in S$ wenn für sie die in Definition 19, Punkt 1b beschriebene Beziehung gilt.*

Im Beispiel gilt $successor(\texttt{he\_has\_}) = \texttt{\_accomplished\_}$.

Definition 24 (followers) *Ein Tupel $f = (f_1, f_2, \ldots, f_m)$ mit allen $f_i \in S$ heißt $followers$ eines Segmentes $s \in S$, wenn*

1. *f_1 successor von s ist.*
2. *für jedes Paar (f_i, f_{i+1}) mit $i < m$ gilt, dass f_{i+1} der successor von f_i ist*
3. *der Endpunkt von f_m gleich n ist, also mit dem Textende zusammenfällt.*

Wir schreiben auch $followers(s) = f$

Die Länge dieses Tupels n bzw. die Zahl der so bezeichneten „Nachfolger“ eines Segmentes heiße auch $|followers(s)|$.

$followers(s)$ ist als Tupel definiert, weil sich die Ordnung der Komponenten eine übersichtliche Definition erlaubt. In intuitiver Notation sprechen wir aber auch von den

Elementen von f, ohne zu beachten, dass es sich um ein Tupel handelt, nicht um eine Menge.

Im Beispiel gilt:

$$followers(\texttt{he\_has\_}) = (\texttt{\_accomplished\_}, \texttt{\_some}) = 2$$

Somit ist $|followers(\texttt{he\_has\_})|$.

Definition 25 (Anfangssegment) *Sei S eine Segmentierung. Dann ist dasjenige $s_{1j} \in S$ mit $j > j'$ für alle $s'_{1j'} \in S$ das Anfangssegment von S.*

In der in Abbildung 2.3b dargestellten Segmentierung ist `he_has_` das Anfangssegment. Analog könnte man ein *Endsegment* definieren.

Technisch realisiert wurde die Implementierung des beschriebenen Formalismus zur Berechnung aller *Segmentierungen* durch ein Perlskript, das auf der in Kapitel 1.2 vorgestellten Suffixbaumimplementierung aufsetzt. In diesem Skript ist die in Abbildung 2.1 dargestellte Matrix als ein Vektor mit Referenzen zu Vektoren repräsentiert.

Das Programm arbeitet rekursiv. Ausgangspunkt ist der Satzanfang. Von dort aus wird Zeichen für Zeichen untersucht, ob der bisherige Textanfang ein *mögliches Segment* ist. Im Erfolgsfall beginnt der Algorithmus aufs neue, mit dem Endpunkt des gefundenen *möglichen Segmentes* als Anfangspunkt. Kann auf diese Weise das Ende des Satzes mit einer Kette von Segmenten erreicht werden, wird der Segmentkandidat als Anfangssegment einer *Segmentierung* gespeichert, ansonsten wird er verworfen. Daraufhin beginnt die Suche von neuem. Für eine detailliertere Beschreibung anhand des bekannten Beispiels siehe Abbildung 2.4.

2.5.2 Die Disambiguierung konkurrierender Segmentierungen

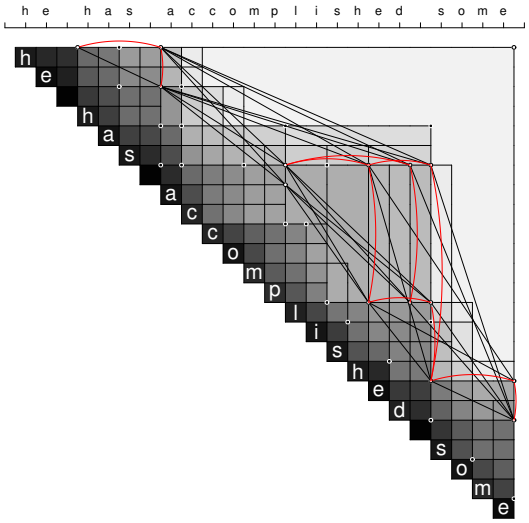
Der Algorithmus wie er sich bisher darstellt, ist alleine noch nicht in der Lage, eine eindeutige Segmentierung zu bestimmen. Die Fülle konkurrierender *Segmentierungen* ist in Abbildung 2.3a dargestellt.

Es bedarf also über den Kernalgorithmus hinaus eines Disambiguierungsverfahrens, das unter den vorgeschlagenen Zerlegungen eine einzige als die beste auswählt.[56]

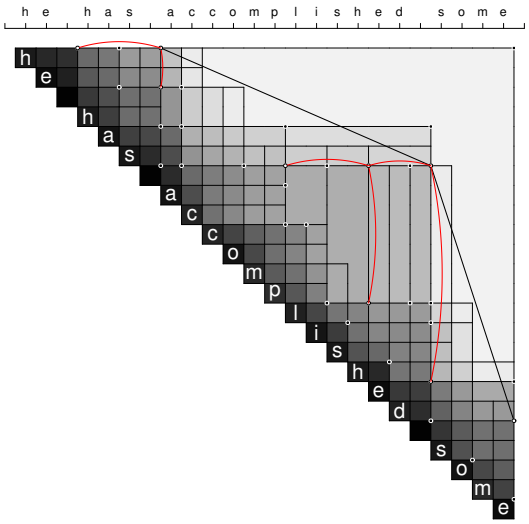
Jeder *Segmentierung* S gemäß Definition 19 wird ein numerischer *Güteindex* $I(S)$ zugewiesen. Je kleiner dieser Index, desto besser wird die entsprechende Segmentierung gewertet. Für die Berechnung von $I(S)$ wurden unterschiedliche Heuristiken vorgestellt und empirisch verglichen. Die verwendete Methode wird im folgenden Abschnitt beschrieben.

Dieses Verfahren eines Rankings mit Hilfe einer Gütefunktion und der empirische Vergleich verschiedener Gütefunktionen ist vergleichbar mit ähnlichen Untersuchungen

[56] Vergleiche auch folgendes Zitat aus Goldsmith (2010, 372): „In general, we may wish to develop an algorithm that assigns a probability distribution over possible analyses, allowing for ranking of analyses: given a string *anicecream*, we may develop an algorithm that prefers *an ice cream* to *a nice cream* by assigning a higher probability to *an ice cream*.“. Diese Stelle spricht sehr ähnliche Probleme wie das vor uns liegende an. Ich würde allerdings wiederum den Begriff „Gütefunktion“ o.ä. gegenüber dem verwendeten „Wahrscheinlichkeitsverteilung“ bevorzugen.



(a)



(b)

Abbildung 2.3: Dieselben Daten in derselben Darstellung wie in den Abbildungen 2.1 und 2.2. In Teilbild (a) sind alle mit Definition 19 kompatiblen Segmentierungen eingetragen, in Teilbild (b) ist nur die vom Algorithmus als optimal bewertete zu sehen. Die zwei Kinder eines *Segmentes* sind mit dem Muttersegment durch rote Kreisbögen verbunden. Die hier dargestellte Segmentierung resultiert aus dem Parametersatz $P_L =$ `combined`, $P_T =$ `tree_sum` und $P_F =$ `none`.

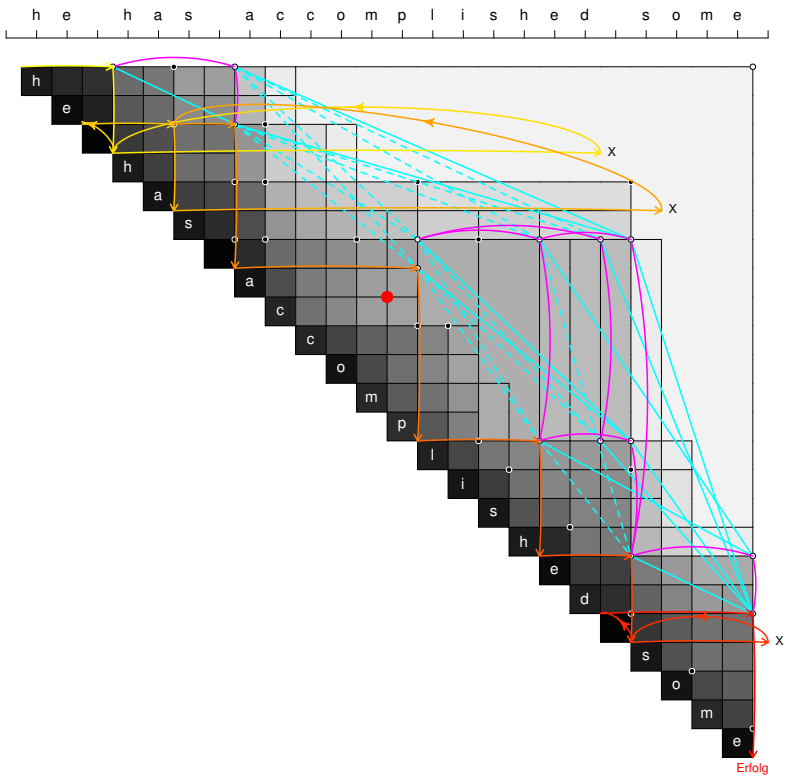


Abbildung 2.4: Der Weg des Algorithmus. Dargestellt sind dieselben Daten wie in den Abbildungen 2.1, 2.2 und 2.3. Grautöne kodieren wieder Häufigkeiten. Die vom Gelben ins Rote wechselnde Pfeilreihe bezeichnet den Weg des Algorithmus. Ausgangspunkt ist der Textanfang in der oberen linken Ecke. Nach rechts gehend wird überprüft, ob ein mögliches Segment vorliegt. `he_` ist ein Treffer. Nun geht es senkrecht nach unten und dann wieder nach rechts, um zu testen, ob es ein unmittelbar anschließendes Segment gibt. In diesem Falle schlägt die Suche fehl. Da das letzte Zeichen des möglichen Anfangssegmentes das Leerzeichen ist, geht der Algorithmus ein Zeichen zurück, um zu testen, ob dort ein mögliches Segment zu finden ist. Mit `_ha` wird er fündig. Hierfür gibt es allerdings wiederum kein mögliches Folgesegment. Da hier auch kein Leerzeichen vorliegt wird die Suche abgebrochen und `_ha` als Segment verworfen. Die Suche für ein Folgesegment von `he_` allerdings wird fortgeführt. `_has_` ist der nächste Kandidat. Von hier aus finden sich im wieder Fortsetzungen. Der dargestellte (Teil)Weg des Algorithmus ergibt die Segmentreihe `he_ _has_ accomp lish ed_ _some`. Alle möglichen Segmente sind durch die Punkte in den Feldecken eingezeichnet. Die weiß gefüllten Punkte sind Teil einer gültigen Segmentierung, die schwarz gefüllten nicht.

von Feng et al. (2004) (vgl. Seite 38). Die Autoren kombinieren für ihre Gütefunktionen Funktionen der Segmentlänge und der *accessor variety*, also der Vielfalt der *möglichen* Fortsetzungen. In meinem Ansatz hingegen wird die Segmentlänge eine viel geringere Rolle spielen und an die Stelle der *accessor variety* tritt u.a. die *Vorhersagbarkeit* wie oben definiert.

Es kommt an manchen Stellen vor, dass der Algorithmus in der bisher beschriebenen Version fehlschlägt. Dies kann trivialerweise geschehen, wenn im Testtext Zeichen vorkommen, die es im Trainingstext nicht giebt. Aber auch Telefonnummern und ähnliche sprachfremde Zeichenketten lassen sich oftmals nicht in mögliche Segmente zerlegen. In einem solchen Fall wird leicht von der vollständigen Überdeckung wie in Definition 19 für Segmentierungen gefordert abgewichen. Dem Algorithmus wird dann erlaubt, erst einzelne Zeichen zu überspringen und dann zur Not auch mehrere. Solche Unterbrechungen in der Segmentierung bezeichne ich als *Brücken*.

Typische Beispiele sind Jahreszahlen wie in folgendem Beispiel aus dem deutschen Testkorpus

```
Feststellen konnte ich, daß um 1625 schon ein Bebel in
Kreuzburg (Schlesien) lebte
```

oder Lehn- und Fremdwörter wie in

```
you're tooling around full of gage in your hot rods, gorging
yourselves on pizza and playing pinball in the taverns and
generally behaving like Ubermenschen.
```

Hier bleiben sowohl das `zz` in `pizza`, als auch das `U` in `Ubermenschen` unsegmentiert zurück.

Abbildung 2.5 zeigt die Häufigkeitsverteilung der Brücken in Abhängigkeit von ihrer Länge. Für alle drei Korpora haben so gut wie alle Brücken nur eine Länge von 1. Der sehr schnelle und exponentielle Abfall der Verteilung zeigt an, dass die nicht segmentierbaren Stellen im Text ein sehr scharf begrenztes und beherrschbares Phänomen darstellen.

Drei Klassen von Heuristiken entscheiden über die Berechnung von $I(S)$. Erstens wird die Güte der Segmente für sich allein bewertet (S. 61). Zweitens werden die entstehenden Segmentbäume als Ganzes bewertet, dh. die längeren Muttersegmente zusammen mit den von ihnen umschlossenen Kindsegmenten (S. 64). Drittens werden die Segmentierungen als ganzes, bestehend aus einer Folge von Segmentbäumen, beurteilt (S. 65).

Die Güte der einzelnen Segmente: P_L

Ein erster Parameter $P_{L(ocal)}$ entscheidet über die Bewertung $I_L(s)$ eines einzelnen, isoliert betrachteten *Segmentes* s. $I_L(s)$ heiße der lokale Güteindex von s.[57] Sechs Möglichkeiten einzelne Segmente zu bewerten wurden getestet:

Sei S eine Zerlegung im Sinne von Definition 19 und $s = s_{ij} \in S$ eines ihrer Segmente. Als Beispiel ziehe ich wieder die Segmentierungen des Testtextes

[57] Ein Ranking von Kandidatensegementen einzuführen, ist an sich kein Alleinstellungsmerkmal meiner Methode. Vergleiche zum Beispiel oben die Diskussion der Arbeit von Cohen et al. (2007).

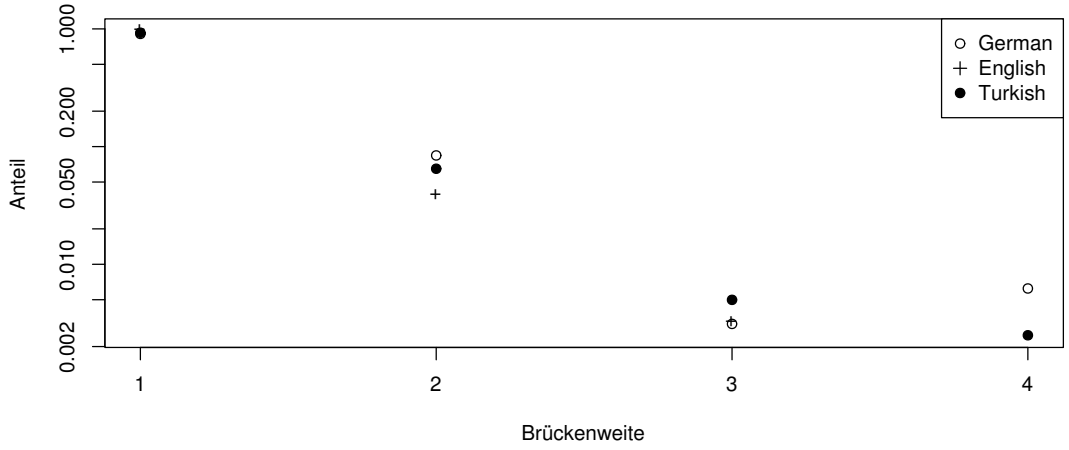


Abbildung 2.5: Verteilung der Brücken in Abhängigkeit von ihrer Breite. Für alle drei Sprachen haben mehr als 90% der Brücken eine Breite von 1. Die Daten wurden anhand leerzeichenfreier und durchgehend kleingeschriebener Testtexte erhoben. (Die Parameterkombination nach der im Verlauf dieses Abschnitts einzuführenden Notation war $P_L =$ `combined`, $P_T =$ `tree_sum`, $P_F =$ `average` und $P_4 =$ `forward_pred`)

`he_has_accomplished_some`

heran. Alle konkurrierenden Segmentierungen sind in Abbildung 2.3a abgebildet.

$P_L =$ `longest` Das längste Segment gilt als das beste:

$$I_L(s) = -l(s)$$

mit der Länge $l(s_{ij}) = j - i + 1$. Man beachte, dass der *Güteindex* $I(S)$ minimiert wird, daher das negative Vorzeichen.

Ganz zu Beginn unseres Beispieltextes gibt es drei mögliche Anfänge zur Segmentierung: $s =$ `he_`, $s' =$ `he_ha` oder $s'' =$ `he_has_`. Falls $P_L =$ `longest`, ergäben sich die Bewertungen $I_L(s) = -3$, $I_L(s') = -5$ und $I_L(s'') = -7$. Die letzte würde als die mit dem niedrigsten I_L ausgewählt.

Die Motivation für die Einbeziehung dieser Heuristik ist, dass das Erkennen möglichst langer Wiederholungen zu hoher Segmentierungsqualität führen könnte.

$P_L =$ `shortest` Das kürzeste Segment gilt als das beste:

$$I_L(s) = +l(s)$$

Falls P_L = `shortest`, ergäben sich die Bewertungen $I_L(s) = 3$, $I_L(s') = 5$ und $I_L(s'') = 7$. Nun würde die erste (s = `he_`) als die mit dem niedrigsten I_L ausgewählt.

Diese Heuristik wurde einbezogen um zu überprüfen, wie weit sich die Strategie der Zerlegung des Inputs in möglichst viele möglichst kurze Segmente auszahlt.

P_L = `forward` Segmente mit einem großen *forward predictability change*, dh. mit $D^+ \ll 1$, werden besser bewertet:

$$I_L(s) = D^+$$

An dieser Stelle wechseln wir im Beispiel zu zwei Segmenten, deren *forward/backward predictability change* wir bereits berechnet haben. Für das oben betrachtete `he_has_` gibt es die möglichen Fortsetzungen `_accomp`, `_accomplish`, `_accomplished` und `_accomplished_`. Wir beschränken uns auf die beiden Segmente, deren Werte wir schon kennen, nämlich `_accomplish` und `_accomplished_`. Es gilt $I_L($`_accomplish`$) = D^+($`_accomplish`$, $`e`$) = 0.51$. Da $I_L($`_accomplished_` $= D^+($`_accomplished_`$, $`s`$)$ mit 0.051 viel kleiner ist, würde das längere Segment hier bevorzugt.

Diese Heuristik ist im Kontrast zur folgenden zu sehen. Es ist eine interessante Frage, ob der Algorithmus bzw. die ihm zugrundeliegende Idee des Vorhersagbarkeitsabfalls in beide Richtungen gleichermaßen funktioniert oder Asymmetrien aufweist.

P_L = `backward` Segmente mit einem großen *backward predictability change*, dh. $D^- \ll 1$, werden besser bewertet:

$$I_L(s) = D^-$$

Hier hatten wir $D^-($`s`$, $`_accomplish`$) = 0.15$ und $D^-($`s`$, $`_accomplished_`$) = 0.22$, womit das kürzere Segment in diesem Falle besser abschneiden würde.

P_L = `combined` Eine Kombination aus `forward` und `backward`:

$$I_L(s) = \log(D^+) + \log(D^-)$$

Bei dieser Einstellung würde sich ergeben:

$$\begin{aligned} I_L(\texttt{\_accomplish}) &= \log(D^+(\texttt{\_accomplish}, \texttt{e})) \\ &\quad + \log(D^-(\texttt{s\_accomplish})) \\ &= \log(0.51) + \log(0.15) = -2.57 \\ I_L(\texttt{\_accomplished\_}) &= \log(D^+(\texttt{\_accomplished\_}, \texttt{s})) \\ &\quad + \log(D^-(\texttt{s\_accomplished\_})) \\ &= \log(0.051) + \log(0.22) = -4.49 \end{aligned}$$

Es würde in diesem Fall das längere Segment als das bessere ausgewählt.

Die in die *Vorhersagbarkeit* eingehenden Substringfrequenzen sind oft von sehr unterschiedlicher Größenordnung. Für den Vergleich drastisch unterschiedlicher Zahlen ist die Betrachtung des Logarithmus oft anschaulicher, da unter der logarithmischen Transformation der Unterschied zwischen 1 und 10 dasselbe Gewicht bekommt wie der Unterschied zwischen 100 und 1000.[58]

$P_L =$ `children` Segmente mit mehr Kindern werden als besser bewertet.

$$I_L(s) = -|leaves(s)|$$

Für $leaves(s)$ siehe Definition 22. In unserem Beispiel wäre $I_L($`_accomplished_`$) = -|leaves($`_accomplished_`$)| = -3$, während $-|leaves($`_accomplish`$)|$ nur -2 ist. Damit würde das längere Segment bevorzugt.

Hinter dieser Heuristik steht eine ähnliche Motivation wie hinter $P_L =$ `shortest`.

Durch P_L wird die Bewertung einzelner Segmente festgelegt. Die nachgeordneten Parameter P_T und P_F steuern nun das Zusammenspiel der Segmente, bzw. die Art und Weise, in der die Bewertungen der Einzelsegmente s in die Bewertung der gesamten *Segmentierung* S eingehen.

Die Güte ganzer Segmentbäume: P_T

In einem zweiten Schritt wird die individuelle Bewertung $I_L(s)$ eines Segmentes s mit den Bewertungen $I_L(s_x)$ und $I_L(s_z)$ seiner Kinder $\{s_x, s_z\} = kids(s)$ zu einem Güteindex $I_T(s)$ zusammengefasst. $I_T(s)$ bewertet nicht nur Segmente, sondern ganze Bäume aus Segmenten. Der Parameter, der diese Berechnung steuert, heißt $P_{T(ree)}$.

Für die Berechnung von I_T wurden drei Möglichkeiten untersucht.

$P_T =$ `tree_none` Bei der Bewertung eines Segments werden mögliche Kinder nicht berücksichtigt:

$$I_T(s) = I_L(s)$$

Für ein Beispiel gehen wir davon aus, dass $P_L =$ `shortest` gesetzt wurde, das heißt $I_L(s) = l(s)$. Betrachten wir wieder die beiden Segmente `_accomplish` und `_accomplished_` im Vergleich, die als *follower* von `he_has_` konkurrieren. Falls nun $P_T =$ `tree_none`, so gilt recht einfach:

$$I_T(\texttt{\_accomplish}) = I_L(\texttt{\_accomplish}) = l(\texttt{\_accomplish}) = 11$$
$$I_T(\texttt{\_accomplished\_}) = 14$$

Damit würde `_accomplish` als besser eingeschätzt.

$P_T =$ `tree_sum` Die Bewertung eines Segments ist die Summe der Bewertungen der Blät-

[58] $\log(10 \cdot a) = \log(10) + \log(a)$

ter im darunterliegenden Baum:

$$I_T(s) = \sum_{l \in leaves(s)} I_T(l) = \sum_{l \in leaves(s)} I_L(l)$$

Die letzte Beziehung gilt, da, falls s keine Kinder hat, also $kids(s) = \emptyset$, immer gilt: $I_T(s) = I_L(s)$, vergleiche Definition 22 (Seite 57).

Im Beispiel würde das bedeuten:

$$\begin{aligned} I_T(\texttt{\_accomplish}) &= I_L(\texttt{\_accomp}) + I_L(\texttt{lish}) \\ &= l(\texttt{\_accomp}) + l(\texttt{lish}) = 7 + 4 = 11 \\ I_T(\texttt{\_accomplished\_}) &= l(\texttt{\_accomp}) + l(\texttt{lish}) + l(\texttt{ed\_}) \\ &= 7 + 4 + 3 = 14 \end{aligned}$$

In diesem Fall ergäbe sich also für P_T = `tree_sum` dasselbe wie für P_T = `tree_none`, dies ist aber für andere Einstellungen von P_L im allgemeinen nicht so.

P_T = `tree_average` Die Bewertung eines Segmentes ist der Durchschnitt der Bewertungen der Blätter im darunterliegenden Baum:

$$I_T(s) = \frac{\sum_{l \in leaves(s)} I_T(l)}{|leaves(s)|} = \frac{\sum_{l \in leaves(s)} I_L(l)}{|leaves(s)|}$$

Im Beispiel ergäbe sich also

$$\begin{aligned} I_T(\texttt{\_accomplish}) &= \frac{I_L(\texttt{\_accomp}) + I_L(\texttt{lish})}{2} \\ &= \frac{l(\texttt{\_accomp}) + l(\texttt{lish})}{2} = \frac{7+4}{2} = 5.5 \\ I_T(\texttt{\_accomplished\_}) &= \frac{l(\texttt{\_accomp}) + l(\texttt{lish}) + l(\texttt{ed\_})}{3} \\ &= \frac{7+4+3}{3} = 4.7 \end{aligned}$$

Bei dieser Einstellung entschiede sich das System also für das längere Segment `_accomplished_`.

Die Güte ganzer Segmentierungen: P_F

Ein dritter Parameter $P_{F(ollower)}$ bestimmt, ob die Güte eines Segmentbaumes isoliert bestimmt wird, oder ob die Bewertung der folgenden Teile des Satzes mit einfließt. Wieder setzt sich der Güteindex, der ein *Segment s* zusammen mit allen seinen Folgesegmenten bewertet, aus den bereits definierten Güteindices zusammen. Auch für P_F wurden drei Möglichkeiten untersucht:

$P_F =$ `none` Jedes *Segment* s wird unabhängig von den folgenden bewertet:

$$I_F(s) = I_T(s)$$

Um ein einfaches Beispiel zu erhalten, setzen wir P_L wieder auf `shortest` und P_T auf `tree_none`. Wie in Abbildung 2.3a zu erkennen, haben unsere beiden Segmente `_accomplish` und `_accomplished_` jeweils genau eine mögliche Fortsetzung.

Auf `_accomplish`folgt `ed_` gefolgt von `_some`
Auf `_accomplished_`folgt `_some`

Die Einzelbewertungen aller auftretenden Segmente sind

$$\begin{aligned} I_T(\texttt{\_accomplish}) = I_L(\texttt{\_accomplish}) &= 11 \\ I_T(\texttt{ed\_}) &= 3 \\ I_T(\texttt{\_some}) &= 5 \\ I_T(\texttt{\_accomplished\_}) &= 14 \end{aligned}$$

Die Beziehung $I_T = I_L$ gilt, da P_T als `tree_none` angenommen wurde.

Falls nun $P_F =$ `none`, so wäre damit auch

$$\begin{aligned} I_F(\texttt{\_accomplish}) = I_T(\texttt{\_accomplish}) &= 11 \\ I_F(\texttt{\_accomplished\_}) &= 14 \end{aligned}$$

Dies, da eventuelle Nachfolgesegmente bei dieser Einstellung nicht berücksichtigt werden. Damit wäre `_accomplish` das bevorzugte Segment.

$P_F =$ `sum` Die Gesamtbewertung eines *Segment(baume)s* s ist die Summe aus seiner eigenen Bewertung und der Bewertungen aller folgenden Segmentbäume:

$$I_F(s) = I_T(s) + \sum_{f \in follower(s)} I_T(f)$$

Für unser Beispiel würde das bedeuten, dass

$$\begin{aligned} I_F(\texttt{\_accomplish}) &= I_T(\texttt{\_accomplish}) + I_T(\texttt{ed\_}) + I_T(\texttt{\_some}) \\ &= 11 + 3 + 5 = 19 \\ I_F(\texttt{\_accomplished\_}) &= I_F(\texttt{\_accomplished\_}) + I_T(\texttt{\_some}) \\ &= 14 + 5 = 19 \end{aligned}$$

In diesem Fall gäbe es also keine eindeutige Entscheidung zwischen beiden Varianten. Welches Segment gewinnt wäre dem Zufall überlassen, in Form der `sort`-Routine von perl. Es ist klar, dass die Kombination aus $P_L =$ `shortest` und $P_F =$ `sum` normalerweise ein unentschiedenes Ranking liefern sollte, da die Länge

bis zum Ende des Satzes von jedem Punkt aus eine Konstante ist. Nur Leerzeichen können eher zufällige Abweichungen hervorrufen, da sie von mehreren Segmenten geteilt werden können. Andere Einstellungen von P_L werden dieses Verhalten nicht zeigen.

$P_F =$ `average` Die globale Bewertung eines Segments ist der Durchschnitt der Bewertungen aller folgenden Segmentbäume.

$$I_F(s) = \frac{I_T(s) + \sum_{f \in follower(s)} I_T(f)}{1 + |follower(s)|}$$

Im Beispiel würde dies bedeuten:

$$\begin{aligned} I_F(\texttt{\_accomplish}) &= \frac{I_T(\texttt{\_accomplish}) + I_T(\texttt{ed\_}) + I_T(\texttt{\_some})}{3} \\ &= \frac{11 + 3 + 5}{3} = 6.33 \\ I_F(\texttt{\_accomplished\_}) &= \frac{I_T(\texttt{\_accomplished\_}) + I_T(\_some)}{2} \\ &= \frac{14 + 5}{2} = 9.5 \end{aligned}$$

In diesem Fall würde sich das System wiederum für `_accomplish` entscheiden.

Nun haben wir alle Mittel beisammen, um in einem letzten Schritt von der Bewertung $I_L(s)$ einzelner Segmente, der Bewertung $I_T(s)$ einzelner Segmentbäume und der Bewertung $I_F(s)$ von Folgen von Segmentbäumen zur Bewertung $I(S)$ der gesamten *Segmentierung* $I(S)$ überzugehen.

$I(S)$ ist definiert als der Index $I_F(a)$ seines Anfangssegmentes a_{ij}. Als Anfangssegment einer Zerlegung gilt nach Definition 25 das längste Segment s_{ij} aus S mit $i = 1$, das also am Anfang des segmentierten Textes steht.

Der Vollständigkeit halber sei hier bereits erwähnt, dass es noch einen vierten Parameter P_4 gibt, der entscheidet, welche Segmente als Kindsegmente verwendet werden, falls hierfür mehrere Möglichkeiten existieren. Der Einfluss dieses Parameters erwies sich als ausgesprochen klein. Deswegen wird er über weite Strecken in der empirischen Evaluation ignoriert werden. P_4 hat 13 mögliche Werte. Um die Darstellung hier nicht zu überfrachten, werden diese erst im Zusammenhang mit Zusammenhang einer eingehenderen Untersuchung des Parameters P_4 beschrieben (Abschnitt 2.6.3).

Insgesamt gibt es also für alle Kombinationen der Parameter $6 \cdot 3 \cdot 3 = 54$ Kombinationen.[59] Soll im Folgenden von einer bestimmten solchen Kombination die Rede sein, so wird hierfür auch der Ausdruck *Parametersatz* verwendet.

In Abschnitt 2.4 (S. 26 ff.) wurde diskutiert, was für Ansätze in der Forschungsliteratur zum Thema untersucht werden. Der aktuelle Abschnitt diente dazu, meinen eigenen Algorithmus vorzustellen. Es ist Zeit für ein paar vergleichende Worte.

[59] Bzw. 702, falls man P_4 miteinbezieht

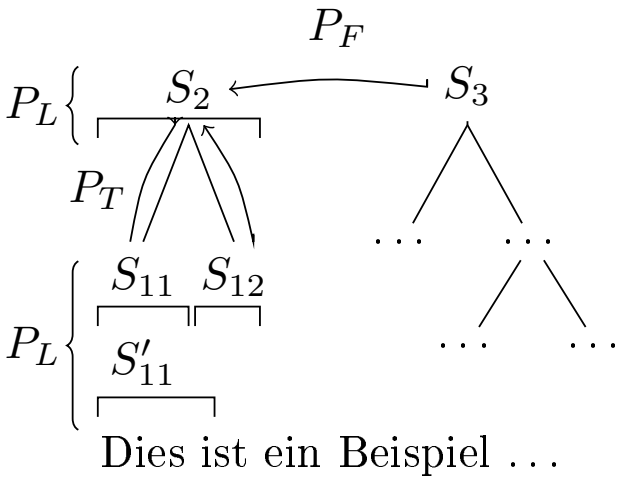


Abbildung 2.6: Bildliche Erläuterung der drei Parameter. P_L bestimmt, nach welchen Kriterien die einzelnen Segmenten bewertet werden. P_T legt fest ob und wie die Kindsegmente bei der Bewertung eines Segmentes berücksichtigt werden. P_F entscheidet darüber ob und wie die nachfolgenden Segmente und deren Bewertung berücksichtigt werden.

Im Gegensatz zu vielen der vorgestellten Arbeiten operiert die hier dargestellte Methode nicht auf Wortlisten oder tokenisiertem Text. Input ist jeweils ein Trainings- und ein Testkorpus als eine einzige Zeichenkette. Es gibt relativ wenige Veröffentlichungen, die hier einen ähnlichen Weg gehen. Zu nennen wäre aber de Marcken (1996) und viele der in 2.4.2 erwähnten Arbeiten wie zum Beispiel Goldwater et al. (2009). Andersherum betrachtet wäre es aber ohne Veränderungen möglich, dem Algorithmus nicht ganze Texten auf einmal als Input zu übergeben, sondern nur einzelne Wörter.

Mit der Praxis der Verarbeitung ganzer Texte ist ein weiterer wichtiger Punkt verknüpft, in der sich meine Arbeit von konkurrierenden Ansätzen unterscheidet: Wenn in der Forschung mit Häufigkeiten von Substrings gearbeitet wird, so im allgemeinen nur bis zu einer vorher festgelegten Länge n. Eine Ausnahme ist mir nicht bekannt. Hier aber gilt diese Einschränkung nicht, Datengrundlage ist die Häufigkeitsstatistik aller im Trainingskorpus vorkommenden Zeichenketten. Dies macht es möglich, den gesamten potentiellen Kontext einer Textstelle auszunutzen.

So wird es möglich, sich nicht auf die lokale Zerlegung von Wörtern oder das lokale Auffinden von Segmentgrenzen zu beschränken, sondern ganze Sätze oder Absätze auf einmal und in ihrem gesamten Zusammenhang zu segmentieren (s. auch Seite 56).

Auf dieser Grundlage wiederum war es nur noch ein kleiner folgerichtiger Schritt, Überlappungen in den entstehenden Zerlegungen zur Gewinnung einer Hierarchie auszunutzen, statt es bei einer rein linearen Segmentierung zu belassen. Auch hier wären mit de Marcken (1996) und den Ansätzen um Goldwater (z.B. Goldwater et al. (2009)) relativ wenige Ansätze zu nennen, die in diesem Punkt in eine ähnliche Richtung gehen.

Viele Arbeiten, vor allem aus der auf Harris (1955) basierenden Klasse leiden erheblich unter dem in Abschnitt 2.4.1 genauer beschriebenen und seitdem mehrfach erwähnten *Abfallproblem*: Nicht nur die Häufigkeiten von Zeichenketten eines Textes nehmen im Allgemeinen mit ihrer Länge (exponentiell) ab. Auch die Zahl der möglichen Fortsetzungen oder die aus ihrer Verteilung abgeschätzte Entropie nehmen tendenziell ab. Abbildung 2.8 zeigt eine kleine empirische Erläuterung der Problematik und einen Vergleich mit dem hier zur Segmentierung verwendeten Maß. Abbildung 2.7 zeigt die Verhältnisse

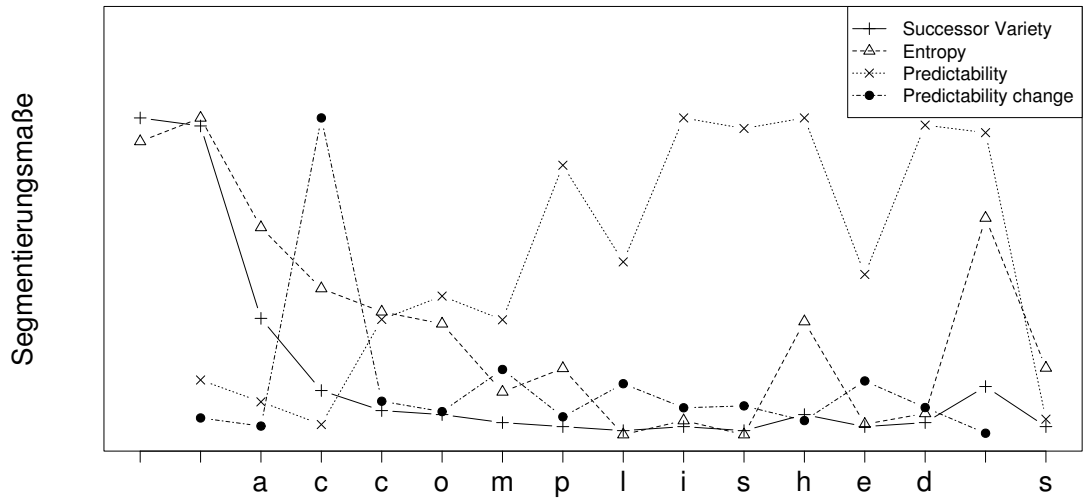


Abbildung 2.7: Entwicklung von *Successor Variety*, *Entropie* und der *Vorhersagbarkeit* und ihrer Änderung für die Zeichenkette `_accomplished_s`. Schon für ein einzelnes Wort zeigt sich, dass die nur der Vorhersagbarkeitsänderung keine abfallende oder ansteigende Grundtendenz hat. Für die y-Achse wurden willkürliche Einheiten gewählt, so dass alle Kurven gut sichtbar sind.

für das `accomplished`-Beispiel. Man kann schließen, dass die traditionellen Maße für sich genommen noch untauglich sind, um *sprachliche Segmente* zu identifizieren. Gewöhnlich werden zusätzliche Heuristiken eingeführt, um dieses Problem zu umgehen.

Auch aus den hier verwendeten vollständigen Substringhäufigkeiten ließe sich keine sinnvolle Zerlegung ableiten, wenn direkt mit relativen Häufigkeiten gearbeitet würde. Aus diesem Grund wird nicht die *Vorhersagbarkeit* der Buchstaben direkt herangezogen, sondern deren Änderung. So wird das Problem nicht umgangen, gemildert oder ausgeglichen, sondern es taucht nicht mehr auf:

Das *Abfallproblem* äußert sich in unserem Beschreibungsrahmen in einer tendenziellen Zunahme der Vorhersagbarkeit für längere Zeichenketten. Ohne Kontext ist der nächste Buchstabe nur sehr schlecht vorhersagbar. Kennt man aber aus einem deutschen Text die Zeichenkette `selbs`, so kann man sich fast sicher sein, dass der nächste Buchstabe ein `t` sein wird. Daher bieten sich genau die Stellen an, an denen dieser allgemeine Trend gebrochen wird und die Vorhersagbarkeit zurückgeht. Abbildungen 2.8 und 2.7 vergleichen die beiden Maße *Vorhersagbarkeit* und *Vorhersagbarkeitsabfall* im Vergleich mit den häufig verwendeten Alternativen *Successor Variety* und *Entropie*.

Abbildung 2.9 ist eine Nebenbemerkung. Hier ist der Anteil der Zeichenketten mit negativem *predictability change* über ihrer Länge aufgetragen. Diese Darstellung beleuchtet die auffällige Entwicklung des *predictability changes* für kleine Stringlängen etwas näher.

Ein weiterer Punkt, der hier noch einmal erwähnt werden soll, sind die minimalen linguistischen Annahmen, die in den Algorithmus einfließen (s. Seite 25 f.). Über den morphologischen Typus oder die Existenz von Suffixen oder Präfixen oder dergleichen wird keinerlei Wissen vorausgesetzt.

Ich habe es explizit vermieden, informationstheoretische Grundlagen für mein Modell entweder vorauszusetzen oder im Nachhinein als Rechtfertigung anzufügen. Solch eine Modellierung müsste meines Erachtens sehr viel rigider gerechtfertigt werden, als dies gemeinhin geschieht. Ohne solche Rechtfertigung verstellen derartige theoretische Annahmen eher den Blick auf die vollen Möglichkeiten, die in den Daten stecken und ziehen tendenziell Folgeprobleme nach sich.

Ein weitere wichtiger Punkt ist im Vergleich zu einem Großteil der bayesianischen Arbeiten anzumerken. Darunter fallen einerseits viele der MDL-Ansätze (2.4.2), als auch die *Hierarchical Bayesian Models* wie von Goldwater und Kollegen (2.4.2) entwickelt. In diesen Arbeiten sind sehr häufig komplizierte oder rechentechnische Suchalgorithmen notwendig um eine globale Kennziffer zu optimieren. Im Kontext der MDL-Arbeiten ist dies zum Beispiel die *Description Length*. Nicht nur kann die Komplexheit solcher Algorithmen ein Problem darstellen. Auch ist es selten möglich, tatsächlich mit Sicherheit das globale Extremum der betrachteten Kenngröße zu finden. Man muss sich in der Regel mit einer Näherung begnügen.

Mein Algorithmus dagegen kommt vollständig ohne einen solchen Suchalgorithmus aus. Im ersten Schritt werden alle mit der *Segment*-Definition im Einklang stehende Segmentierungen ermittelt. Im zweiten Schritt wird nach einem eindeutigen Rankingverfahren eine daraus ausgewählt. Dies kontrastiert stark mit den ausgeklügelten Minimierungs- bzw. Maximierungsverfahren, die in den erwähnten Arbeiten eingesetzt werden.

Als Fernziel ist anvisiert, zu untersuchen, ob es Möglichkeiten gibt, die sehr unterschiedlichen Ansätze wie sie die Gruppe um Goldwater und ich vorschlagen, konstruktiv miteinander zu verbinden.

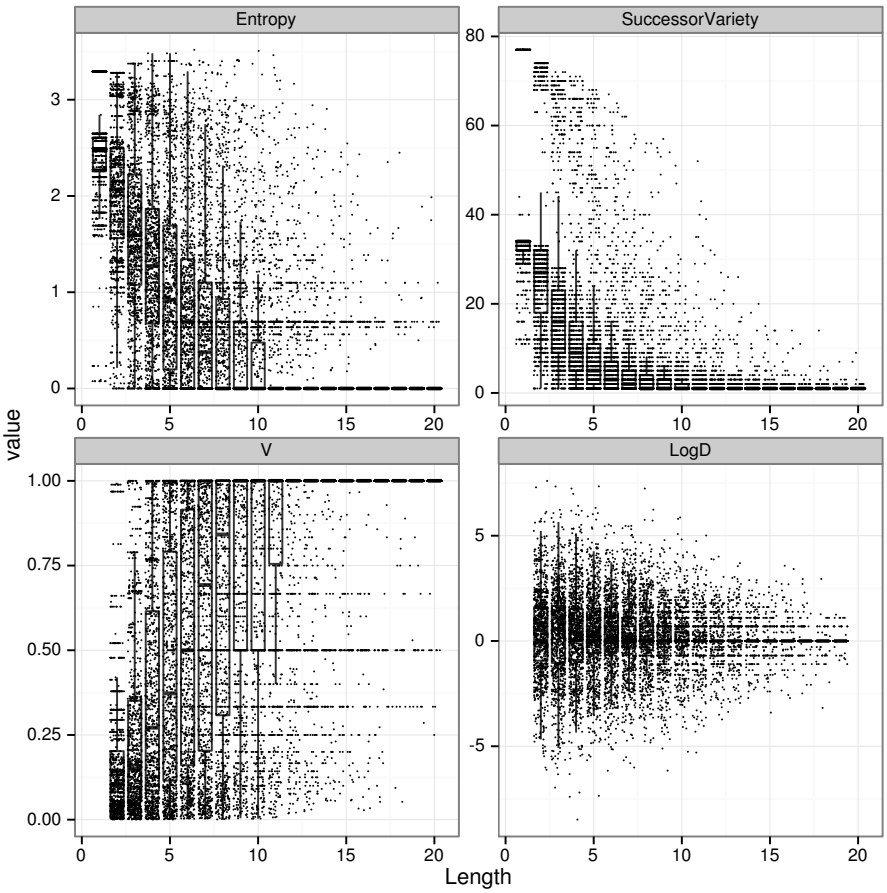


Abbildung 2.8: Die in der Literatur zur Segmentierung verwendeten Maße *Successor Variety* und *Entropie* im Vergleich zu Vorhersagbarkeit (V) und Vorhersagbarkeitsabfall ($\log(D)$). Auf der x-Achse ist jeweils die Länge eines zufällig ausgewählten Strings dargestellt. Auf der y-Achse das entsprechende Segmentierungsmaß. *Entropie* und *Successor Variety* fallen mit der Stringlänge ab. Man erkennt in beiden Verteilungen eine Häufung von Punkten oberhalb der Masse der Werte. Diese Untermenge sollte die Kandidaten für Segmente darstellen. Auch ihre Verteilung fällt allerdings ab, so dass jeder Cutoff zumindest von der Stringlänge abhängen sollte, ein Vorgehen, dass mir so aus keinem Artikel bekannt ist. Die Vorhersagbarkeit steigt mit der Stringlänge stark an. Nur der Vorhersagbarkeitsabfall ist für längere Strings klar um Null zentriert. Das *Abfallproblem* existiert hier nicht. Für kleine Längen existiert eine klare Struktur: Die Kurve beginnt oberhalb von Null, steigt noch etwas und fällt dann auf Null ab. Diese Struktur könnte sich als interessant erweisen. Alle Daten wurden am Browncorpus (Francis und Kucera, 1967) erhoben.

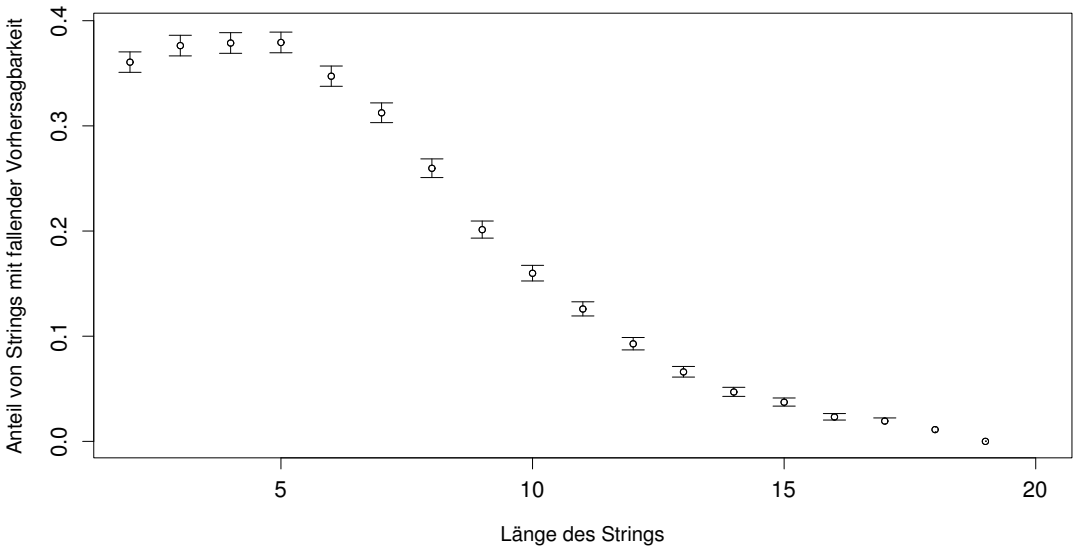


Abbildung 2.9: Anteil der Strings s mit Vorhersagbarkeitsabfall ($D^+(s) < 1$) aufgetragen über der Länge von s. Auffällig ist der oberhalb einer Länge von 5 beginnende Abfall der Kurve.

2.6 Empirische Evaluation des Algorithmus

In diesem Abschnitt soll es darum gehen, den vorgeschlagenen Algorithmus zu evaluieren und seine durch $P_{L,F,T,4}$ definierten Varianten zu vergleichen.

Dabei ist ein Ziel, den vorgestellten Algorithmus als unüberwachte Segmentierungsmethode zu etablieren. Darüberhinaus ist es an verschiedenen Punkten möglich, aus Aspekten im Verhalten der verschiedenen Verfahrensvarianten linguistisch relevante Schlüsse zu ziehen.

Evaluation kann hier kaum etwas anderes heißen als ein Vergleich der Segmentierungsvorschläge des Systems mit einer linguistischen Analyse. Für diesen Vergleich wird die linguistische Analyse gewöhnlich in einen *Goldstandard* umgesetzt, also einen Text, für den die gestellte Aufgabe bereits gelöst wurde, meist durch menschliche Experten (siehe Seite 25). Die Auswertung erfolgt dann üblicherweise mit Hilfe sogenannter *Evaluationsmaße*, z.B. *Recall*, *Precision* und *f Measure*, die wir in den kommenden Abschnitten noch näher kennenlernen werden.

Ein derartiges Vorgehen ist etabliert. Ein prominentes Beispiel auf dem Gebiet der *Morphologischen Induktion* ist der bereits erwähnte *Morphochallenge*[60] (Kurimo et al., 2010). Dennoch ist das Verfahren nicht ohne Probleme. Für den vorliegenden Fall lassen sich zwei Hauptschwierigkeiten ausmachen:

Das erste Problem ist spezifisch für das hier untersuchte Verfahren: Meines Wissens gibt es keinen öffentlich verfügbaren Goldstandard, der das vom Algorithmus gelieferte Format reproduziert. Dazu bedürfte es theoretisch einer Baumbank bis hinunter auf die morphologische Ebene. Ein solches Goldstandard-Korpus für die drei untersuchten Sprachen zu erstellen wäre ausgesprochen aufwendig. Der Gewinn würde sich dagegen in relativ engen Grenzen halten, da es keinerlei Grundlage für einen Vergleich mit konkurrierenden Algorithmen gäbe. Nicht einmal ein Vergleichskandidat ist mir bekannt.[61]

Verfügbare Goldstandards bestehen aus einem Text, in dem die Segmentgrenzen gekennzeichnet sind. Auch ein solches Korpus ist nicht ohne Probleme.

Das zweite Problem, das ich erwähnen möchte ist grundsätzlicher Natur und betrifft alle relevanten Ansätze gleichermaßen. In Abschnitt 2.2 wurde ein Begriffsgerüst morphologischer Einheiten entworfen. Dort ist charakterisiert, was im Rahmen dieser Arbeit gemeint sein soll, wenn von „Segment“, „Morph“, „Morphem“ und ähnlichen Entitäten die Rede ist. Leider reicht das noch lange nicht aus, in jedem konkreten Fall zu entscheiden, wo eine Segmentgrenze zu setzen ist und wo nicht. Verschiedene Linguisten mit unterschiedlicher theoretischer Ausrichtung werden unterschiedliche Antworten geben auf die Frage, ob eine Wortform wie „Ursache“ aus einem oder zwei Morphen besteht. Ein Linguist, der Gewicht auf den diachronen Charakter von Sprache legt wird eine solche Frage vielleicht bejahen, da das Wort aus den zwei Konstituenten *Ur-* und *Sache* gebildet wurde (Grimm und Grimm, 1984, 2502). Ein Linguist mit Hauptaugenmerk auf dem derzeitigen System der deutschen Sprache wird wohl verneinen, da das Wort

[60] Website des 2010'er Wettbewerbs: `http://research.ics.tkk.fi/events/morphochallenge2010/`

[61] Obwohl die Zahlen nicht streng und vor allem nicht systematisch quantitativ vergleichbar sind, fallen alle vergleichenden Betrachtungen zwischen meinen Ergebnisse und den von anderen Forschern publizierten Performanzwerten positiv aus.

im modernen Deutsch als Simplex wahrgenommen wird. In vielen Fällen sind darüber hinaus auch noch weitere verschiedene Sichtweisen möglich.

In Abschnitt 2.6.3 wird empirisch gezeigt, dass diese Theorieabhängigkeit mindestens 15% der vorstellbaren Segmentgrenzen betrifft. Dies ist meines Erachtens eine Größenordnung, die eine fundierte Auseinandersetzung mit dem Thema verlangt. In der in Abschnitt 2.6.3 vorgestellten Untersuchung werde ich daher die angesprochene Variabilität direkt in die Evaluation einbeziehen.

Die Tatsache, dass es kein allgemeingültiges Evaluierungskorpus geben kann, da die Theorie keine unumstrittene Segmentierung bereitstellt, ist aber nur eines von mehreren Problemen.

Im Falle des naheliegendsten Kandidaten für einen nutzbaren Goldstandard, des Morphochallenge-Referenzkorpus, kommt hinzu, dass es nicht manuell erstellt wurde, sondern auf automatischen Verfahren beruht. Diese variieren sogar zwischen den enthaltenen Sprachen. Die tatsächliche Qualität des Korpus und die Verlässlichkeit einzelner Segmentgrenzen ist daher unbekannt.

Dennoch könnte man argumentieren, dass der Gewinn an Vergleichbarkeit bei Verwendung dieses Korpus vergleichsweise groß wäre, da es bereits in der Realität mehrfach als Goldstandard eingesetzt wurde. Aber auch dieses Korpus ist in der Community nicht als allgemeiner Evaluierungsstandard akzeptiert, vgl. auch Seite 28.

Den angesprochenen Problemen begegne ich mit drei verschiedenen Untersuchungen: Leerzeichen bilden in den Schriftsystemen der drei behandelten Sprachen Deutsch, Englisch und Türkisch so gut wie immer auch Segmentgrenzen. Diese Untermenge an Segmentgrenzen ist unbestritten und theorieunabhängig. So ergibt sich ein leicht zu gewinnender partieller Goldstandard. Eine daran orientierte Evaluation wird in Abschnitt 2.6.2 durchgeführt.

Nur diese Segmentgrenzenuntermenge zu untersuchen wäre unbefriedigend. Daher wird in Abschnitt 2.6.3 doch ein kleiner, extra erstellter Goldstandard verwendet um den Algorithmus genauer zu evaluieren. Dieser trägt der unvermeidlichen Theorieabhängigkeit Rechnung, indem die Unstimmigkeiten verschiedener Experten explizit berücksichtigt werden.

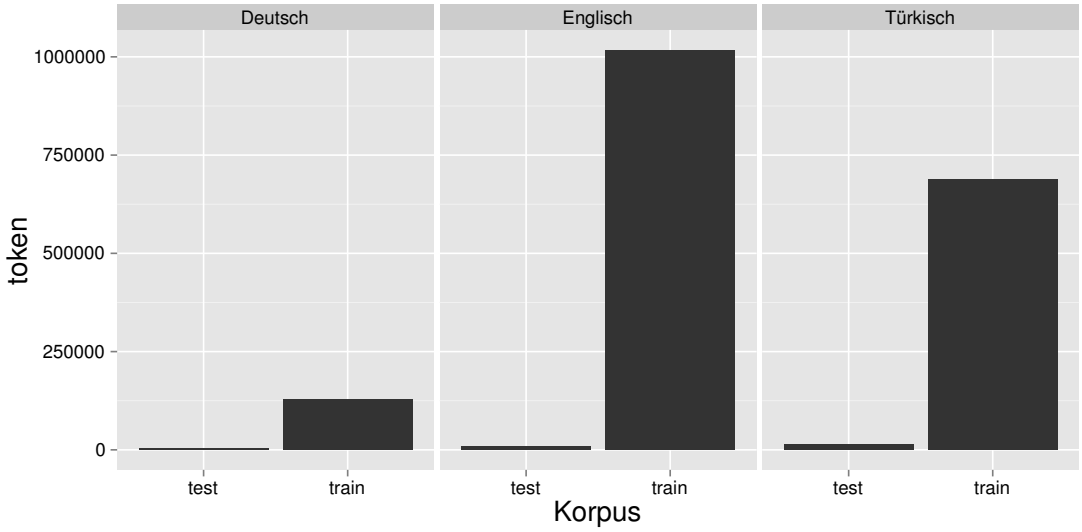
Aber auch dies wäre noch nicht vollständig. Schließlich segmentiert der Algorithmus nicht nur auf Morphemebene, sondern ordnet auch kleinere Segmente zu größeren zusammen. Bisher wurden allerdings nur die Grenzen zwischen Segmenten untersucht, nicht ganze Segmente und erst recht nicht Segmente, die ihrerseits wieder aus anderen Segmenten bestehen. In Abschnitt 2.6.4 wird daher untersucht, was für einen linguistischen Status die vom Algorithmus vorgeschlagenen Segmente haben. Da ein passendes Goldkorpus wie erwähnt fehlt und das Kostennutzenverhältnis einer eigenen Erstellung nicht überzeugend wäre, beschränke ich mich hier auf die manuelle Inspektion einer relativ kleinen Stichprobe aus den vollen Daten. Ein Hauptaugenmerk wird hierbei auf der Klassifikation der Fehler liegen, die das System macht.

Dabei ergibt sich ein klares Bild, von dem vermutet werden kann, dass es nicht nur für den vorliegenden Ansatz von Bedeutung ist, sondern für eine große Klasse verwandter Ansätze ebenfalls Gültigkeit haben dürfte. Zum Abschluss des Kapitels skizziere ich einen zweiten, ergänzenden Mechanismus, mit dem kombiniert der vorgestellte Algorithmus

vielleicht zu einem wesentlich mächtigeren Werkzeug der automatischen Textanalyse werden könnte.

2.6.1 Die verwendeten Daten

Datengrundlage der experimentellen Untersuchungen waren deutsche, englische und türkische Texte. Tabelle 2.1 bietet einen Überblick über die verwendeten Korpora.



Sprache	Quelle	Trainingskorpus	Testkorpus
Deutsch	Bebel (2004a,b)	$1.28 \cdot 10^5 tk$	$4.41 \cdot 10^3 tk$
		$9.01 \cdot 10^5 chr$	$2.94 \cdot 10^4 chr$
			$200l$
Englisch	Kuĉera und Francis (1967)	$1.01 \cdot 10^6 tk$	$9.03 \cdot 10^3 tk$
		$5.89 \cdot 10^6 chr$	$5.12 \cdot 10^4 chr$
			$500l$
Türkisch	Say et al. (2002)	$6.89 \cdot 10^5 tk$	$1.45 \cdot 10^4 tk$
		$5.44 \cdot 10^6 chr$	$1.10 \cdot 10^5 chr$
			$410l$

Tabelle 2.1: Verwendete Trainings- und Testkorpora. Größenangaben in *Token* (tk), *Zeichen* (chr) und Zeilen (l). Die begleitende Graphik verdeutlicht die Zahlenangaben (Token).

Die Trainingsdaten werden jeweils verwendet, um die notwendigen Substringfrequenzen $N_T(s)$ zu bestimmen. Auf deren Grundlage berechnen sich die *forward/backward predictability changes* $D_T^{\pm}$ der Zeichenketten des Testkorpus. Daraus wiederum folgen

die möglichen *Segmentierungen* des Textes. Diese werden (ebenfalls mit Hilfe der Trainingsfrequenzen) disambiguiert, nach Maßgabe der Parameterwerte für $P_{L,T,F,4}$.

Der Text wird nicht nur in seiner Originalform analysiert, sondern in insgesamt 4 Versionen, die sich aus der Kreuzung zweier Variablen ergeben:

representation Enthält der Text Leerzeichen? Es ist hier entscheidend, dass der Algorithmus in keinem Fall von vornherein über das Wissen verfügt, dass Leerzeichen Segmentgrenzen sind. Die einzige im Algorithmus vorhandene Sonderregel ist, dass Leerzeichen zu zwei aufeinanderfolgenden Segmenten gleichermaßen gehören können. Dies gilt für kein anderes Zeichen.

- `no` Der Text enthält Leerzeichen.[62]
- `s` Der Text enthält keine Leerzeichen mehr. Die Entfernung der Leerzeichen ermöglicht es zu überprüfen wie hilfreich die Leerzeichen für die Berechnung der Segmentgrenzen ist und wie sich das Fehlen der Leerzeichen auf die verschiedenen Aspekte des Verfahrens auswirkt.

case Wurde die Groß- und Kleinschreibung normiert?

- `uc` Groß- und Kleinschreibung wurde in der originalen Version belassen.
- `lc` Alle Großbuchstaben sind auf Kleinbuchstaben normiert. Diese Operation hat erwartungsgemäß vor allem für das Deutsche einen merklichen Einfluss, da nur hier intensive Großschreibung existiert.

Die Texte wurden folgendermaßen vorbereitet: Häufungen von Leerzeichen, Tabulatoren und Zeilenumbrüchen wurden vor der Verarbeitung jeweils aus den Test- und Trainingstexten auf ein einzelnes Leerzeichen reduziert. Zeilenumbrüche tauchen also nicht weiter auf und müssen nicht berücksichtigt werden. Diese Operation wurde durchgeführt, damit zwei aufeinanderfolgende Wörter immer nur genau durch ein Leerzeichen getrennt sind. Dies kann die Statistik für wortübergreifende Zeichenketten ein wenig verbessern.

2.6.2 Vollständige Evaluation der Rückgewinnung von Leerzeichen

Dieser Abschnitt hat zwei Ziele: Zum einen soll die allgemeine Tragfähigkeit des Ansatzes anhand eines verlässlichen Datensatzes gezeigt werden. Zum anderen wird bereits der Einfluss der in 2.5.2 beschriebenen Parameter auf die Performanz des Algorithmus untersucht.

Obwohl der Algorithmus viel weitergehende Daten zu produzieren in der Lage ist, beschränken sich die folgenden Überlegungen bis Abschnitt 2.6.4 auf die Evaluation der gefundenen Segmentgrenzen.

Nehmen wir für den Augenblick an, die Menge der sprachlichen Segmente eines Textes wäre vollständig bekannt. Dann könnte man aus dem Vergleich der maschinellen Segmentierung mit diesem Goldstandard die Evaluationsmaße *Recall* und *Precision* bestimmen.

[62] Die Benennung orientiert sich daran, was *entfernt* wurde. In diesem Fall also nichts.

Diese sind zwar recht weitläufig bekannt. Da ich sie in einer der folgenden Untersuchungen aber modifizieren werde (Abschnitt 2.6.3), füge ich ihre explizite Definition hier noch einmal ein.

Der *Recall* beschreibt, wie viele der im Goldstandard existierenden Segmentgrenzen als solche erkannt werden:

Definition 26 (Recall) *Die Zahl N_c der korrekt identifizierten Grenzen sprachlicher Segmente geteilt durch die Zahl N_G der im Goldstandard existierenden derartigen Grenzen bildet den Recall:*

$$R = \frac{N_C}{N_G}$$

Die *Precision* bewertet die Performanz von der anderen Seite her und gibt an, wie viele der vorgeschlagenen Segmentgrenzen korrekt sind:

Definition 27 (Precision) *N_C geteilt durch die Zahl N_S der in der maschinellen Segmentierung enthaltenen Segmentgrenzen bildet die Precision:*

$$Pr = \frac{N_C}{N_S}$$

Nun ist die vollständige Zerlegung eines Textes in *minimale sprachliche Segmente* und erst recht deren Zusammenordnung zu kompletten Analysebäumen übergreifender *sprachlicher Segmente* – wie im vorigen Abschnitt dargestellt – hochgradig theorieabhängig und für die untersuchten Korpora schlicht nicht vorhanden.

Es sei daran erinnert, dass die Leerzeichen für den Algorithmus zwar Zeichen einer speziellen Art sind, aber nicht von vornherein als Morphemgrenzen betrachtet werden. Der Grundgedanke der folgenden Untersuchung ist daher folgender: Man kann wohl durchaus davon ausgehen, dass in jedem Schriftsystem, das überhaupt Leerzeichen zur Worttrennung verwendet, folgende Regel gilt: Einem Leer- oder Satzzeichen entspricht immer[63] auch die Grenze eines *sprachlichen Segmentes* nach Definition 6. Etwas unscharf könnte man auch sagen, eine „Morphemgrenze“. So bekommen wir eine Untermenge von Segmentgrenzen, die sich eindeutig identifizieren lassen. Dies erlaubt für diese Untermenge die Angabe eines *Recall*. Da die vollständige Zerlegung weiterhin unbestimmt bleibt, kann aber nach wie vor nichts der *Precision* vergleichbares angegeben werden.

Da unser „Goldstandard“ zu diesem Zeitpunkt nicht vollständig ist und somit der *Recall* streng genommen nicht existiert, definiere ich eine dem *Recall* ähnliche Hilfsgröße, die *Performanz*.

Definition 28 (Performanz) *Sei n_c die Zahl der vom Algorithmus gesetzten Segmentgrenzen, die mit einem Leerzeichen zusammenfallen. Sei n_s die Zahl der tatsächlich oder ursprünglich im Testtext(abschnitt) vorhandenen Leerzeichen. Dann ist die Performanz des Algorithmus definiert als*

$$P = \frac{n_c}{n_s} \tag{2.7}$$

[63] Spezielle Ausnahmen wie gesperrte Schreibweise können hier ignoriert werden. Meiner Einschätzung nach können sie kaum ins Gewicht fallen.

P bezeichnet also den Anteil der „gefundenen“ Leerzeichen. Wenn man davon ausgehen könnte, dass sich für die Grenzen zwischen *sprachlichen Segmenten*, die nicht mit Leerzeichen zusammenfallen, jeweils derselbe P-Wert ergäbe, so wäre dieser gleich dem *Recall* für die volle Menge der Morphemgrenzen.

Die grundlegende Annahme hinter den folgenden Überlegungen und Experimenten ist, dass ein Parametersatz, der besser als andere in der Lage ist, die Grenzen zwischen orthographischen Wörtern in einem Text zu finden, auch die allgemeinere Aufgabe der Segmentierung in *sprachliche Segmente* besser lösen kann. Sonst wäre die nachfolgende Untersuchung von begrenztem Wert.

Im Allgemeinen ist diese Annahme schwer zu überprüfen. Für ein abschließendes Urteil wäre es erforderlich, sehr viele Segmente händisch zu evaluieren. Im darauf folgenden Abschnitt (2.6.3) wird eine Überprüfung durch den Vergleich mit einem (kleinen) Goldstandard dennoch direkt durchgeführt, mit erstaunlich klaren und positiven Ergebnissen.

Um die verschiedenen Parametereinstellungen untereinander vergleichen zu können, wird das Testkorpus jeweils in kurze Abschnitte zerlegt: Sätze für Englisch und Deutsch, Paragraphen für Türkisch[64]. Für jeden dieser Abschnitte wird berechnet, wie viele der ursprünglich im Text enthaltenen Leerzeichen mit einer *Segmentgrenze* der vom Algorithmus ausgegebenen *Segmentierungen* zusammenfallen.

Ich fasse die variierten Parameter noch einmal kurz zusammen:

- Zum einen gibt es die drei untersuchten Sprachen: Deutsch (deu), Englisch (eng) und Türkisch (tur) (S. 75). Streng genommen wird die Sprache nicht als Parameter behandelt, sondern die Untersuchungen finden jeweils getrennt nur für eine Sprache statt.
- P_L mit 6 möglichen Werten (S. 61 ff.).
- P_F (S. 65 ff.) und P_T (S. 65 ff.) mit jeweils drei möglichen Werten.
- *representation* mit zwei möglichen Werten (S. 76).
- *case* ebenfalls mit zwei möglichen Werten (S. 76).

Insgesamt werden die 200–500 Testsätze/abschnitte der drei Korpora in jeweils allen $3 \cdot 3 \cdot 6 \cdot 2 \cdot 2 = 216$ Parameterkonstellationen untersucht.

Wie in den Abbildungen 2.14 bis 2.16 zu erkennen und auf Seite 2.6.2 diskutiert, haben die Parameter untereinander im Allgemeinen starke Wechselwirkungen. Dh., der Einfluss des einen Parameters auf die Performanz hängt selbst wieder vom Wert der anderen Parameter ab. Daher ist es keine leichte Aufgabe, die in den Daten vorhandenen Strukturen zu erkennen und zu beschreiben. Ich beginne mit der Beschreibung der wichtigsten summarischen Kennzahlen. Anschließend werden einfacher zu erfassende Zusammenhänge mit deskriptiver Statistik, das heißt mithilfe von Graphiken, beleuchtet. Ab einer gewissen Komplexität werden Graphiken zwangsläufig unübersichtlich. Über diesen Punkt gehe ich mit einem halb heuristischen, halb analytischen Ansatz hinaus.

[64] Dies hat technische Gründe.

Bei diesen Untersuchungen sind sowohl Gemeinsamkeiten, als auch Unterschiede zwischen den Sprachen interessant. Vorhersagbare Gemeinsamkeiten dienen als Konsistenztest und zeigen, dass sich der Algorithmus erwartungsgemäß und linguistisch sinnvoll verhält. Auch manche der Unterschiede zwischen den einzelnen Sprachen reflektieren zum Teil bekannte Ähnlichkeits- bzw. Unähnlichkeitsverhältnisse zwischen Deutsch, Englisch und Türkisch. Darüber hinaus gibt es weitere, klar erkennbare, aber nicht so leicht ad hoc zu deutende Unterschiede zwischen den verschiedenen Sprachen. Diese deuten an, dass ein Potential von neuen und linguistisch möglicherweise nicht uninteressanten Strukturen in den Daten verborgen liegen könnte.

Abbildung 2.10 gibt einen ersten Überblick über die maximale Performanz in den drei Sprachen (Abbildung 2.10). Diese schwankt in absoluten Zahlen relativ stark, vor allem zwischen Deutsch und Englisch auf der einen Seite und Türkisch auf der anderen. Insgesamt ist dieses erste Ergebnis ausgesprochen ermutigend. Zum einen sind die erre-

Sprache	Maximale mittlere Performanz
Deutsch	0.961
Englisch	0.951
Türkisch	0.888

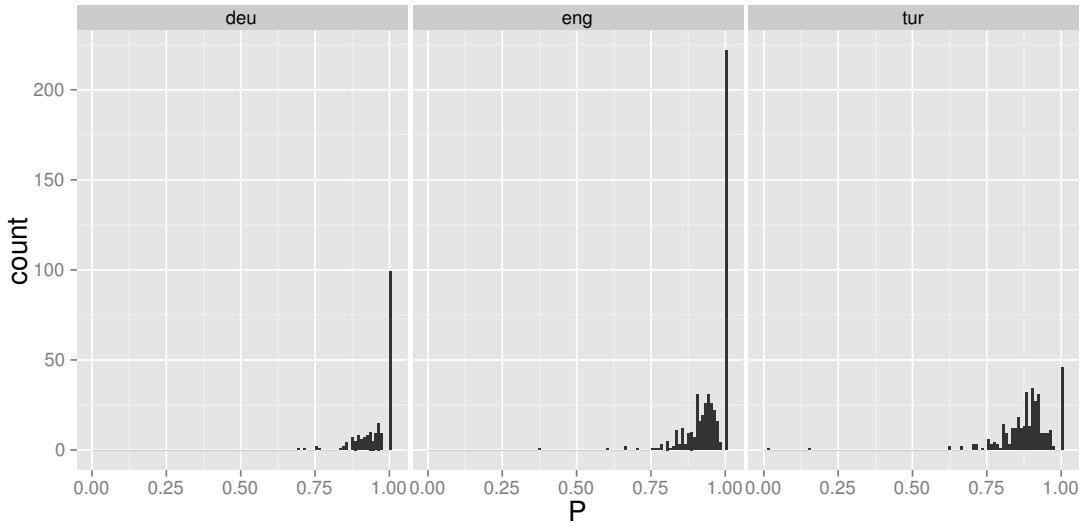


Abbildung 2.10: Die Performanzverteilung für alle Sätze bei optimalen Parameterwerten für die untersuchten Sprachen. Dargestellt sind die Häufigkeiten über den Performanzwerten P. Dies bedeutet beispielsweise, dass in 100 der 200 deutschen Testsätze alle Leerzeichen als Segmentgrenzen erkannt wurden. Der optimale Parametersatz ist sprachunabhängig: $P_L =$ `combined`, $P_T =$ `tree_sum`, $P_F =$ `sum`, *representation* = `no` und *case* = `uc`.

ichten Zahlen vor allem für Deutsch und Englisch mit einer Quote von nur $4-5\%$ nicht als *Segmentgrenzen* erkannten Leerzeichen auf den ersten Blick sehr brauchbar, auch

wenn eine abschließende Bewertung oder gar ein Vergleich mit anderen Ansätzen an dieser Stelle nicht möglich ist. Die Mittelwerte geben allerdings nur ein unvollkommenes Bild über die Performanz des Algorithmus. In der beigefügten Graphik ist die Verteilung der P-Werte der einzelnen Sätze explizit dargestellt. Man erkennt vor allem im Englischen den hohen Anteil an Sätzen mit $P = 1$. Der Verdacht liegt nahe, dass es sich hier um triviale Fälle handeln könnte, zum Beispiel extrem kurze Sätze. Dies ist nicht der Fall, der Median der Leerzeichenzahl für die Untermenge der Sätze mit $P = 1$ liegt in allen drei Sprachen weit über 1.

Zum anderen ist der optimale Parametersatz in allen drei Sprachen identisch ($P_L =$ `combined`, $P_T =$ `tree_sum`, $P_F =$ `sum`, *representation* = `no` und *case* = `uc`). Stellt sich dies als eine stabile Eigenschaft des Algorithmus heraus, so wäre das für jede mögliche Anwendung ein glücklicher Umstand, da der Algorithmus damit sprachunabhängig einsetzbar wird.

Auffällig ist, dass von den untersuchten Korpora das türkische Korpus deutlich die schlechteste *Performanz* erzielt.[65] Es ist nicht von vornherein auszuschließen, dass dieser Performanznachteil des Türkischen auf tiefere, linguistisch relevante Ursachen zurückgeht. So ist es denkbar, dass der vorgestellte Segmentierungsalgorithmus für türkische Texte ganz allgemein nicht so geeignet ist wie für deutsche oder englische. Prinzipiell vorstellbar sind sprachtypologische Gründe.

Im Laufe der Untersuchung zeigten sich aber viele auffällige parallele Strukturen in den Daten über alle drei Sprachen hinweg. Dies lässt die Vermutung naheliegend erscheinen, dass es keinen qualitativen Unterschied der Performanz des Algorithmus zwischen Türkisch und den westgermanischen Sprachen gibt.

Es gibt zwei wahrscheinlichere Erklärungsansätze für den Performanzabfall im Türkischen. Einerseits wären Wechselwirkungen mit Topic und Genre zu untersuchen. Das türkische Korpus besteht wesentlich aus innenpolitischen Artikeln. Diese enthalten viele Abkürzungen und Eigennamen. Auch unentdeckte Verunreinigungen des türkischen Korpus scheinen nicht unmöglich (S. Fußnote 65). Für eine abschließende Klärung der Frage wird es notwendig sein, alternative türkische Korpora zu analysieren.

In einem nächsten Schritt behandle ich den Einfluss der beiden Parameter, die nicht die Feineinstellung des Algorithmus selbst betreffen, sondern die spezifische Form der Daten: *representation* mit den beiden Werten `no` (mit Leerzeichen) und `s` (ohne Leerzeichen) und *case*, das die beiden Werte `uc` („upper case“) und `lc` („lower case“) annehmen kann.

Den Effect von *case* zu untersuchen dient zwei Zwecke: Einerseits lassen sich aus der Struktur der Orthographie der drei untersuchten Sprachen einfache Vorhersagen ableiten, deren empirische Überprüfung eine Plausibilitätsprüfung der Ergebnisse (*proof*

[65] Zu Beginn der Untersuchungen schien der Unterschied sogar nicht unwesentlich größer zu sein. Eine Durchsicht der Daten ergab, dass ein gewisser Teil des Testtextes unanalysiert blieb. Die letztendliche Ursache waren Wiederholungen ganzer (Ab)-Sätze im Korpus. Da der Algorithmus auf der vollständigen Statistik aller Wiederholungen im Text basiert, reagiert er empfindlich auf derart unnatürliche Wiederholungen. Die Struktur der Daten allerdings erlaubt es auch, diese wiederholt auftretenden Sätze zu erkennen und von der Analyse auszuschließen. Die problematischen Sätze sind eindeutig daran zu erkennen, dass sich der Satz als ganzes oder über 90% bereits im Trainingstext finden. Dies geht aus den verwendeten Daten unmittelbar hervor. Eine ausschnittsweise Sichtung der Texte zeigte keine weiteren Auffälligkeiten.

of concept) ermöglicht und die ausreichende Auflösung der Analysemethoden belegt. Andererseits lassen sich aus den Resultaten linguistische Hypothesen ableiten, deren weitere Überprüfung fruchtbar sein könnte.

Die Regeln für Groß- und Kleinschreibung im Englischen und Türkischen gleichen sich im wesentlichen. Außer Wörtern am Satzanfang und Eigennamen wird grundsätzlich klein geschrieben. Daher kann vorhergesagt werden, dass sich die beiden Sprachen in Bezug auf *case* in ihrem Verhalten nicht wesentlich unterscheiden. Im Deutschen werden dagegen zusätzlich alle Nomina groß geschrieben. Dies führt in etwa zu einer Verdopplung des Anteils der Großbuchstaben am Text.[66]

Aus der grundsätzlichen Gleichheit des Systems im Englischen und Türkischen kann man die Hypothese ableiten, dass der Einfluss der Variable *case* in diesen beiden Sprachen sehr ähnlich sein sollte. Die Abweichende Position des Deutschen hingegen legt es nahe, dass *case* hier unterschiedliche Wirkung hat.

Die Konvertierung aller Großbuchstaben in Kleinbuchstaben in einem Text kann für die hier untersuchte Anwendung im allgemeinen zwei gegenläufige Folgen haben, je nachdem, ob Eigennamen oder Substantive betroffen sind. Eigennamen bestehen häufig aus Zeichenketten, die in keiner anderen Funktion auftreten wie `Garland`, `Reyhan` oder `Friedrich`. Hier ist die Großschreibung ein wichtiger Hinweis auf die Wortart. Nivelliert man die Großschreibung, so wird Information aus dem Text entfernt. Falls das überhaupt einen messbaren Effekt hat, sollte es der Performanz des Segmentierungsalgorithmus eher schaden. Werden Substantive systematisch groß geschrieben, sind sie von diesem Informationsverlust ebenfalls betroffen. Daneben gibt es aber einen zweiten Effekt. Die groß geschriebenen Anfangsmorpheme vieler deutscher Substantive erscheinen kleingeschrieben in anderen Wortarten (`Leben`, `lebst`) oder wortintern (`Bildung`, `Vorbildung`). Hier hat die Nivellierung der Großschreibung zur Folge, dass diese Morpheme in allen Kontexten gleich realisiert werden und daher vom Algorithmus auch als gleich identifiziert werden können. Die Verbesserung der Statistik für die einzelnen Morpheme sollte sich positiv auswirken.

Diese Überlegungen führen zu weiteren Hypothesen in Bezug auf den Einfluss des Parameters *case* auf die Performanz des untersuchten Segmentierungsalgorithmus. Da im Englischen und Türkischen (im Wesentlichen) nur Eigennamen groß geschrieben werden, sollte die Nivellierung der Großschreibung keinen positiven, sondern allenfalls einen negativen Einfluss haben. Im Deutschen hingegen existieren beide Effekte, der negative durch den Informationsverlust in Bezug auf die Wortart, und der positive durch eine Verbesserung der Statistik für identische Morpheme. Ob die Nivellierung der Großschreibung hier insgesamt einen positiven oder negativen Effekt hat, ist also im vorhinein nicht zu entscheiden. Falls sich eine unterschiedliche Auswirkung nachweisen lässt, sollte der Einfluss auf die Performanz im Deutnschen positiver sein als im Englischen und Türkischen.

In den alphabetischen Schriftsystemen ist die deutsche Orthographie eine Ausnahme. Vom deutschen Sprachraum aus breitete sie sich kurzzeitig in Teile Skandinaviens aus,

[66] Im englischen Trainingskorpus gibt es 2.0% Großbuchstaben, während es im deutschen Trainingskorpus 4.4% sind.

konnte sich aber dort nicht halten.[67] Wenn die eben dargelegten Hypothesen sich durch die Daten stützen lassen, wäre es naheliegend zu untersuchen, ob sich die Erkenntnisse auf die Lesbarkeit eines Textes für Menschen übertragen lässt. Das wäre wiederum ein Argument für die These, dass das Deutsche mit seinem Orthographiesystem nicht umsonst alleine ist, sondern weil es einen ungünstigen Sonderweg genommen hat.

Einen ersten Überblick über die Daten gibt Abbildung 2.11 . Der Unterschied zwis-

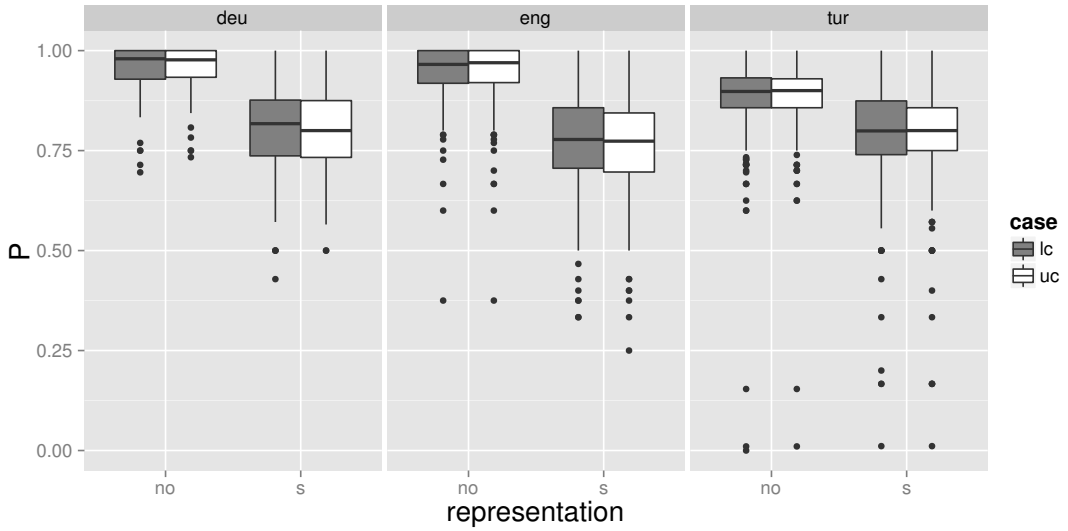


Abbildung 2.11: Graphische Darstellung des Einflusses der beiden Parameter *representation* und *case* auf die Performanz des Algorithmus. Die übrigen Parameter sind auf die optimalen Werte (P_L = `combined`, P_F = `sum`, P_T = `tree_sum`) festgeschrieben. Auf der x-Achse sind die Repräsentationen aufgetragen. `no` steht für den Originaltext, `s` für die Textversion ohne Leerzeichen. Die y-Achse zeigt den Anteil der vom Algorithmus gefundenen Leerzeichen P an.

chen auf Kleinschreibung normalisierten Korpora (`lc`) und der Originalschreibweise (`uc`) scheint insgesamt nicht besonders groß.

Der erste Blick auf die Daten bestätigt die aufgestellten Hypothesen nur eingeschränkt: In den englischen und türkischen Daten finden sich kaum Unterschiede zwischen `uc` und `lc`. Auch im Deutschen gibt es nur in der leerzeichenfreien `s`-Version einen Vorteil für die normalisierte Schreibweise: Hier liegt der Median bei 0.82 und damit 2 Prozentpunkte über der unnormalisierten Darstellung `uc`. Dieser Unterschied erscheint zwar recht gering, im Verhältnis zu den fünf übrigen Fällen ist er aber nicht unerheblich. Dieser erste Überblick über die Daten stützt also nur Teile der aufgestellten Hypothesen: Englisch und Türkisch verhalten sich sehr ähnlich, während das Deutsche abweicht. Was die Richtung der Unterschiede angeht, werden die Hypothesen durch Abbildung 2.11 nicht gestützt.

[67] Das Norwegische (vor 1869) und das Dänische (vor 1948) (Wikipedia-Mitarbeiter, 2005) folgten zeitweise der deutschen Großschreibung.

In Abschnitt 2.6.2 (Seite 88) und in Abschnitt 2.6.3 (Seite 108) ergeben sich bei genaurer Betrachtung der Daten weitere Aspekte zu dieser Problematik.

Ob das Korpus Leerzeichen enthält (`no`) oder nicht (`s`) wirkt sich dagegen deutlich auf die Performanz aus. In Abbildung 2.11 ist dies klar zu erkennen. Dort ist zwar nur der Teil der Daten mit optimaler Parametereinstellung dargestellt. Aber auch über alle übrigen Parameterwerte gemittelt fällt die Performanz um 15% (Deutsch), bzw. 16% (Englisch). Im Türkischen ist der Abfall mit nur 9% deutlich kleiner. Dieser geringere Unterschied im Türkischen passt sehr gut zu der Tatsache, dass diese Sprache von einer außergewöhnlichen morphologischen Regelmäßigkeit ist. Beispielsweise gibt es im Türkischen nur zwei unregelmäßige Verben, *yemek* (essen) und *demek* (sagen), und auch deren Unregelmäßigkeit äußert sich lediglich in gelegentlichen Anpassungen des Stammvokals. Eine ähnlich strenge Regelmäßigkeit lässt sich auch in der türkischen Morphologie als ganzes beobachten. Diese Eigenschaft ist vor allem für Altaische Sprachen, allgemeiner aber auch für agglutinierende Sprachen insgesamt typisch.[68] Eine solch hohe Stabilität der Morpheme lässt eine bessere Segmentierbarkeit auch ohne Leerzeichen natürlich erscheinen.

Diesen Unterschied zwischen den beiden westgermanischen Sprachen und Türkisch auf der derzeit verfügbaren Faktengrundlage allzu hoch zu bewerten zu wollen, wäre aber müßig. Im Endeffekt ergeben sich in der `lc`-Version für alle drei Sprachen sehr ähnliche Performanzwerte.

Dass die *Performanz* für die Textrepräsentation mit Leerzeichen höher ist als für den Text ohne Leerzeichen kann auch damit zu tun haben, dass der Algorithmus ganz allgemein mehr Segmentgrenzen setzt, wenn es Leerzeichen gibt: Dies rührt daher, dass Leerzeichen ungleich allen anderen Zeichen zu zwei benachbarten Segmenten gehören dürfen (siehe Definition 19 und die vorausgehende Diskussion).

Es wäre daher denkbar, dass die höhere Performanz für Repräsentationen mit Leerzeichen alleine auf die zufällige Verteilung der zusätzlich gesetzten Segmentgrenzen zurückzuführen ist. An dieser Stelle macht es sich leider bemerkbar, dass die dargestellte Evaluationsmethode nur eine dem *Recall* ähnliche Größe liefert, ohne etwas über die *Precision* auszusagen. Diese Lücke wird erst in 2.6.3 geschlossen.

case und *representation* beschreiben die Form der Daten. Wenden wir uns nun dem Einfluss der drei algorithmusinternen Parameter P_L, P_T und P_F zu.

Es wäre vorstellbar, dass die drei Parameter unabhängig voneinander auf die Performanz des Algorithmus einwirken. Dies würde bedeuten, dass eine spezielle Stellung eines bestimmten Parameters denselben positiven oder negativen Effekt hat, unabhängig von Veränderungen in den übrigen Parametern. In einem solchen Fall spricht man von additivem Verhalten oder einfachen Haupteffekten. Verändert sich jedoch der Effekt, den ein Parameter hat, in Abhängigkeit von weiteren Parametern, so wird diese Wechselwirkung als Interaktion bezeichnet.

Der gleichzeitige Einfluss sämtlicher Parameter auf die mittlere Performanz lässt sich im Prinzip über graphische Darstellungen der Mittelwerte für alle Parameter verbildlichen wie in Abbildung 2.12. Der starke gegenseitige Einfluss der Parameter

[68] Amir Zeldes, pers. Komm.

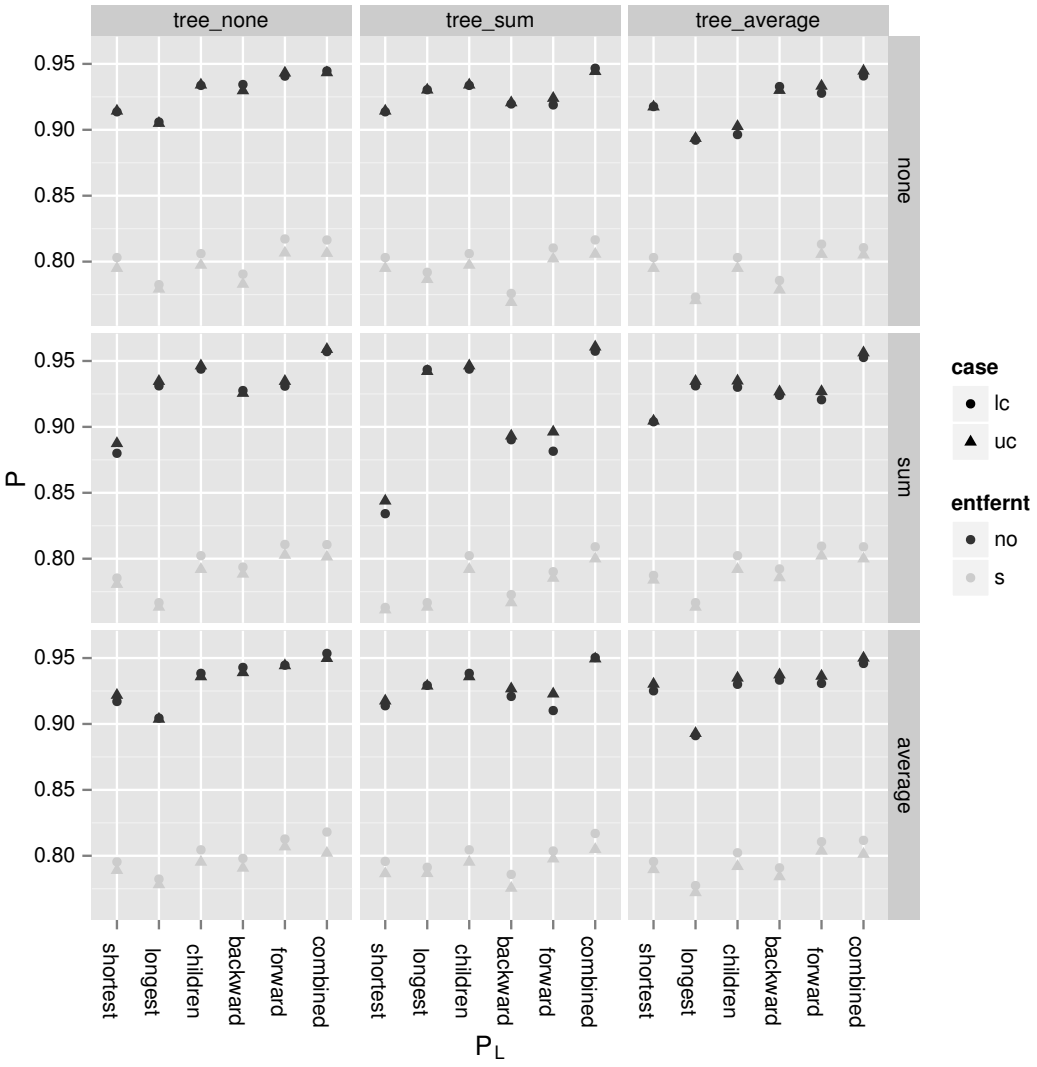


Abbildung 2.12: Die *Performanz* (P) des Algorithmus für alle Kombinationen der Parameter $P_{L,F,T}$, *representation* und *case*. Dargestellt sind die deutschen Daten. Die X-Achse zeigt die 6 möglichen Werte für P_L, auf der Y-Achse sind die arithmetischen Mittelwerte der *Performanz* über alle Sätze dargestellt. Die Zeilen der Einzelfelder zeigen die 3 möglichen Werte für P_F, die Spalten die drei P_T-Werte.

aufeinander ist hier klar erkennbar. So zeigt zum Beispiel das mittlere Teilbild von 2.12, dass sich die sechs Parametereinstellungen für P_L für P_F = `sum` und P_T = `tree_sum` deutlich anders verhalten als die für die übrigen Werte von P_F und P_T. Es liegen also offensichtlich Interaktionen vor, die in der Evaluation auch berücksichtigt werden müssen.

Um nun eine Ordnung in das Zusammenspiel dieser Wechselwirkungen zu bekommen

und um abschätzen zu können, welche Eigenschaften sich in allen drei Sprachen finden und welche sprachspezifisch sind, reicht der rein deskriptive Blick auf die Daten allerdings nicht aus. Das Bild der einzelnen Teildaten ist zu komplex, als dass man es zu einem Gesamteindruck vereinigen könnte.

Daher ist dies der Punkt, an dem über die rein deskriptive Statistik mit Bildern hinausgegangen werden muss. Ich gehe im Folgenden erst auf die von mir verwendete Methodik ein. Anschließend komme ich auf den Einfluss der drei Parameter $P_{L,T,F}$ zurück.

Die naheliegende Grundidee ist, statistische Modelle aufzustellen, die alle 5 Parameter und ihre Wechselwirkungen bis zur ersten Ordnung beinhalten.

Als Responsevariable bzw. als abhängige Variable bietet sich die *Performanz* an. Diese berechnet sich nach Definition 28 aus den Zahlen der vorhandenen und gefundenen Leerzeichen in einem Satz. Typischerweise sind derartige Größen, die die Zahl von „Erfolgen“ in einer Anzahl von „Versuchen“ beschreiben, binomialverteilt. Gewöhnlich ist die Binomialverteilung exzellent durch eine Normalverteilung beschreibbar. Dies lässt an eine Modellierung mittels des allgemeinen linearen Modells denken, bzw. gewöhnlicher Regression. Deren Anwendbarkeit hat aber drei Voraussetzungen: Die Unabhängigkeit der einzelnen Datenpunkte, Normalverteilung der *Residuen* und die Unabhängigkeit der *Residuenvarianz* von den Parametern. Als Residuen werden die Schwankungen der Messwerte um den vom Modell vorhergesagten Erwartungswert bezeichnet. Abbildung 2.13 verdeutlicht den Begriff.

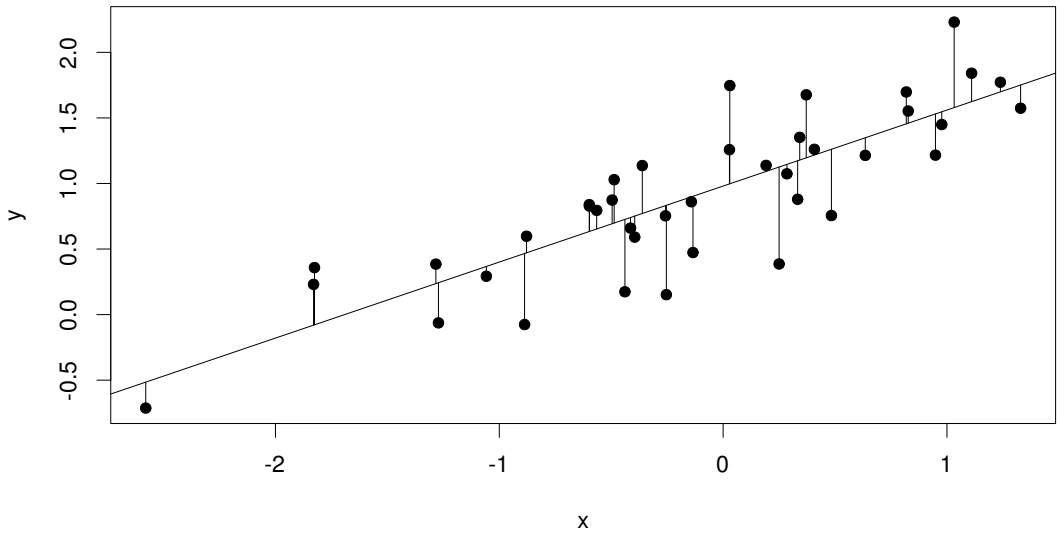


Abbildung 2.13: Graphische Verdeutlichung des Begriffs der Residuen. Die ausgefüllten Punkte sind die Datenpunkte eines hypothetischen Experimentes. Die durchgezogene schräge Linie repräsentiert das am besten passende Modell für den tatsächlichen Zusammenhang zwischen x und y (lineare Regression). Die senkrechten Verbindungslinien zwischen Messungen und Modellvorhersagen bilden die Residuen.

Nun ist es aber leider so, dass die Normalverteilungsnäherung der Binomialverteilung in den Randbereichen der Verteilung, nahe 0 oder 1, zusammenbricht. Abbildung 2.10 zeigt deutlich, dass wir uns in diesem Randbereich befinden. Außerdem hängt die Varianz der Binomialverteilung[69] vom Erwartungswert ab. Bei 0 und 1 ist diese Abhängigkeit besonders stark. Beide Eigenschaften machen die Anwendung gewöhnlicher Regressionsmethodik unangemessen.

Einen Ausweg bieten die generalisierten Modelle (siehe unter anderem Pinheiro und Bates (2000); Baayen (2008); Zuur et al. (2009)). Diese machen es möglich, über die Modellklasse normalverteilter Residuen hinauszugehen und direkt binomialverteilte Daten zu modellieren. Daten sind genau dann binomialverteilt, wenn in einer Serie von „Versuchen" die Zahl der „Treffer" gezählt wird und sich die Wahrscheinlichkeit für einen Treffer durch eine konstante Wahrscheinlichkeit p charakterisieren lässt. Ein Standardbeispiel ist die Zahl der Sechser in einer Serie aus n Würfelwürfen. Übertragen auf die vorliegende Situation bedeutet dies, dass jede Segmentgrenze einen Versuch repräsentiert. Mit einer Wahrscheinlichkeit p findet der Algorithmus sie, mit der Wahrscheinlichkeit $1 - p$ übersieht er sie. Jeder Satz ist in dieser Darstellung eine Serie von Versuchen und die Zahl der gefundenen Segmentgrenzen binomialverteilt.

Innerhalb der generalisierten Modelle gibt es eine Unterklasse, die auch berücksichtigt, dass manche Sätze aufgrund zufälliger Eigenschaften schwieriger zu segmentieren sind als andere. Die *gemischten generalisierten Modelle* erlauben, dass die Wahrscheinlichkeit p, mit der eine Segmentgrenze erkannt wird, von Satz zu Satz (normalverteilt) schwankt. Die Varianz dieser Schwankung geht als Parameter in das Modell ein. Der Name *gemischte Modelle* oder *mixed models* gründet sich darin, dass sowohl Varianzen solch *zufälliger* Effekte Parameter des Modells sind, als auch der genaue Einfluss spezifischer Parameterstellungen *fester* (*fixed*) Variablen. Die Parameter $P_{L,T,F,4}$ gehen als solche *festen Effekte* in das Modell ein. Auf Seite 102 ergibt sich eine ähnliche Diskussion in einer vergleichbaren Situation.

Aber auch dieses Modell hat unleugbare Schwierigkeiten. Es ist nicht anzunehmen, dass die Zahl der erfolgreich gesetzten Morphemgrenzen wirklich streng binomialverteilt ist. Dies würde voraussetzen, dass die Wahrscheinlichkeit, dass ein bestimmtes Leerzeichen als Segmentgrenze gesetzt wird, innerhalb eines Satzes von Leerzeichen zu Leerzeichen konstant ist. Statt dessen wird es wiederum leicht und schwer zu entdeckende Segmentgrenzen geben. Dies widerspräche der Annahme einer Binomialverteilung: Wenn die Wahrscheinlichkeit für eine Sechs nicht konstant ist, so ist die Zahl der gewürfelten Sechser nicht mehr binomialverteilt.

Auch diese Situation ließe sich modellieren, indem man eine Ebene tiefer ansetzt und als abhängige Variable modelliert, ob ein bestimmtes Leerzeichen als Segmentgrenze gesetzt wurde, oder nicht. Nun gäbe es nicht einfach oder schwer zu segmentierende Sätze, sondern einfach oder schwer auffindbare Leerzeichen.

Dies würde aber wiederum voraussetzen, dass die Wahrscheinlichkeiten, jeweils aufeinander folgende Leerzeichen als Segmentgrenzen zu identifizieren, unabhängig

[69] $np(1-p)$; n die Zahl der Versuche, p die Wahrscheinlichkeit für einen „Erfolg", bzw der Erwartungswert der Verteilung.

voneinander sind. Dies ist wahrscheinlich nicht der Fall. Stattdessen wird ein (Miss)erfolg bei einem Leerzeichen die Wahrscheinlichkeit für einen (Miss)erfolg beim folgenden Leerzeichen beeinflussen. Damit gilt die Unabhängigkeit der Datenpunkte nicht mehr. Es gibt zwar auch Möglichkeiten derartige Komplikationen in generalisierte Modelle einzubauen. Unter anderem die Implementierung von Pinheiro et al. (2011) im Rahmen des `R`-Paketes `nlme` lässt dies zu. Dies würde bei sorgfältiger Analyse vielleicht erlauben, die Struktur und Stärke der wechselseitigen Abhängigkeiten genauer einzugrenzen. Diese Korrelationsstruktur zu untersuchen könnte durchaus ein lohnendes Ziel zukünftiger Forschung sein.

Die Erfahrung zeigt jedoch, dass ein Vernachlässigen derartiger Korrelationsstrukturen qualitativ meist nicht viel an den übrigen ermittelten Abhängigkeiten und Effekten verändert, auf denen hier das Hauptaugenmerk liegt. Ein möglicher Einfluss ist allerdings ein Aufblähen der berechneten p-Werte (s. z.B. Zuur et al. (2009, S. 350)). Dies ist im vorliegenden Fall aber unproblematisch, da die p-Werte ohnehin nur von begrenztem Wert sind.

Entgegen den meisten Anwendungsfällen für statistische Modelle gibt es hier nicht das Problem mangelnder Daten. Im Gegenteil lassen sich mit ein wenig Rechenzeit beliebig viele Daten generieren. Daraus folgen zwei Probleme, eines begrifflicher und eines rechentechnischer Natur. Das rechentechnische Problem ist trivial: Die riesige Menge der zur Verfügung stehenden Datenpunkte führt schnell zu unpraktikablen Rechenzeiten.

Das begriffliche Problem ist altbekannt und hier wegen der beliebigen Verfügbarkeit von Daten ungewöhnlich offen sichtbar. In aller Regel sind Nullhypothesen über die Gleichheit von Populationsgrößen formuliert. Ein Beispiel wäre: „Die Wahrscheinlichkeit, eine Sechs zu würfeln, ist 1/6.“. Für einen realen Würfel wird dies niemals in letzter Konsequenz zutreffen, da immer kleine Materialungleichmäßigkeiten existieren. In diesem Sinne ist jede Nullhypothese streng genommen immer falsch.

Je mehr Daten vorliegen, desto feinere Unterschiede werden messbar. Wenn die Stichprobe nur groß genug ist, wird sich immer eine signifikante Abweichung von der Nullhypothese finden lassen. Man kann argumentieren, dass damit der Begriff der Signifikanz selbst wertlos wird, der ja dazu dienen soll, die Einfluss nehmenden Variablen zu identifizieren (vergleiche zu dieser Problematik z.B. Vicente und Torenvliet (2000); Cohen (1994)).

Da im vorliegenden Fall die Datenmenge beliebig vergrößert werden kann, tritt dieses Problem hier besonders deutlich zu Tage. Daher betrachte ich direkt die Schätzwerte für die einzelnen Parameter anstelle von p-Werten. Um dennoch einen realistischen Eindruck von deren Stabilität zu bekommen, werden die Modelle jeweils auf 5 unterschiedlichen Samples der Daten berechnet. Jedes Sample bestand aus $5 \cdot 10^4$ Datenpunkten.

Damit ist das angekündigte technische Problem ebenfalls umgangen: Die schnell anwachsende Rechenzeit, die benötigt wird, um die entsprechenden Modelle auf den vollen zur Verfügung stehenden Daten auszuwerten.

Ich fasse das verwendete Modell noch einmal kurz zusammen. Die Zahl der als Segmentgrenzen identifizierten Leerzeichen wird als binomialverteilt angenommen. Die Wahrscheinlichkeit, ein Leerzeichen als Segmentgrenze zu identifizieren ist innerhalb eines Satzes eine Konstante und zwischen den Sätzen normalverteilt. Innerhalb eines

Satzes werden die Erfolgswahrscheinlichkeiten als unabhängig modelliert.

Insgesamt hat das verwendete Modell seine Schwächen, dh. es bildet von vornherein nicht alle Aspekte der Daten ab. Dies ist der komplexen Natur der Daten geschuldet. Es ist aber mit keinem großen Einfluss auf die qualitativen und quantitativen Voraussagen des Modells zu rechnen. Darüber hinaus kann angenommen werden, dass etwaige Verzerrungen nicht nur in einem der drei Korpora (Deutsch, Englisch, Türkisch) auftreten, sondern in allen in ähnlichem Ausmaß zum Tragen kommen. Insofern kann aus einem Vergleich der Modellparameter in den verschiedenen Sprachen durchaus gültige Schlüsse gezogen werden. Zur Berechnung der Modelle wurde das `R`-Paket `lme4` (Bates et al., 2011) eingesetzt.

Die graphisch dargestellten Ergebnisse finden sich in den Abbildungen 2.14 bis 2.16. Dort bedeuten positive (negative) Werte auf der x-Achse, dass ein Parameterwert einen positiven (negativen) Einfluss hat. Der *Intercept* repräsentiert die Performanz bei der Parameterstellung $P_L =$ `shortest`, $P_T =$ `tree_none`, $P_F =$ `none`, *representation* = `no`, *case* = `lc`. Weitere Details wie diese Graphiken zu lesen sind, finden sich in den Bildunterschriften. Hier kann man noch einmal deutlich sehen, dass die Interaktionen zwischen den Parametern eine wesentliche Rolle spielen: Andernfalls wären die Haupteffekte weiter von Null entfernt als die Interaktionen. Dies ist nicht der Fall, dh. die Wechselwirkungen zwischen den Parametern sind von derselben Größenordnung wie ihre Haupteffekte. Dies verkompliziert die Deutung der Haupteffekte. Spielen zwei Parameter zu stark zusammen, ist es schwer, ihre Wirkung jeweils für sich zu betrachten. Die Haupteffekte beschreiben allein nicht mehr die volle Wirkung eines Parameters, sondern nur noch seinen Einfluss gemittelt über den übrigen Parameterraum. Behält man dies im Gedächtnis, kann die Betrachtung der Haupteffekte dennoch aufschlussreich sein. Ich beginne mit ihrer Diskussion.[70]

Der Parameter *representation***:** Aus der auf Abbildung 2.11 folgenden Diskussion erwarten wir, dass der Parameter *representation* einen starken Einfluss besitzt, genauer, dass *representation* = `s` wesentlich schlechter abschneidet als *representation* = `no`. Dies erkennt man in den Abbildungen 2.14 bis 2.16 auch tatsächlich daran, dass *representation* = `s` mit einem stark negativen Wert in allen drei Sprachen die letzte oder vorletzte Zeile belegt.

Der Parameter *case***:** Die Auswirkung des Parameters *case* scheint in der Übersichtsgraphik Abbildung 2.11 (Abschnitt 2.6.2, Seite 82) recht gering. Die Verlässlichkeit dieser Aussage allein aufgrund der graphischen Darstellung der Verhältnisse bei ansonsten fixierten Parametern $P_{L,T,F}$ ist allerdings relativ gering. Die hier präsentierten statistischen Modelle stellen ein wesentlich mächtigeres Analysemittel dar.

Ich rekapituliere die in 2.6.2 entwickelten Hypothesen: Aus der Kenntnis der Orthographie der untersuchten Sprachen heraus kann erwartet werden, dass sich Englisch und

[70] Eine Nebenbemerkung zu den türkischen Daten in Abbildung 2.16: Die Schwankungen der Ergebnisse sind hier stärker als für Deutsch und Englisch, obwohl sie auf derselben Zahl an Datenpunkten beruhen. Dies könnte mit den auf Seite 80 beschriebenen Problemen mit diesem Korpus zusammenhängen.

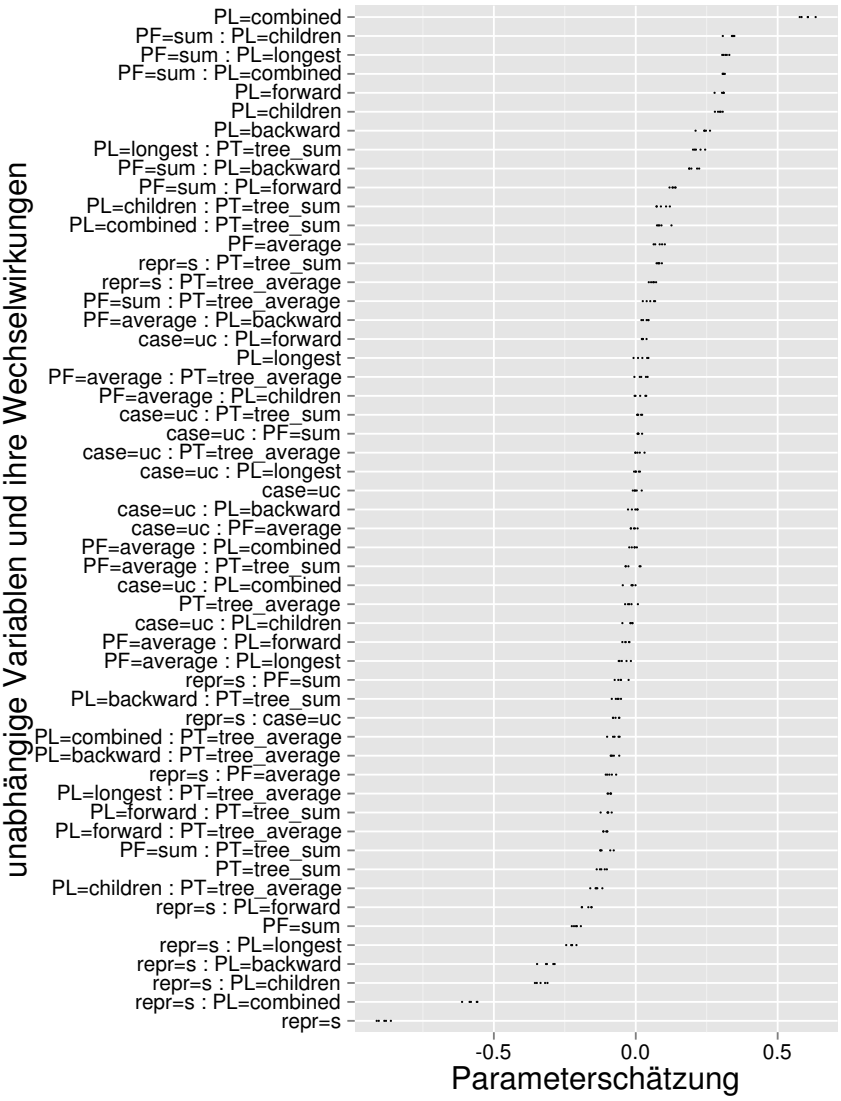


Abbildung 2.14: Die *deutschen* Daten. Übersicht über die Parameterwerte, die sich für das im Text beschriebene gemischte generalisierte Modell ergeben. Positive/negative Werte auf der X-Achse korrespondieren mit einem positiven/negativen Einfluss der entsprechenden Parameterstellungen. Die Streuung in x-Richtung repräsentiert die Schwankung über die 5 Fitdurchläufe. Die Parameter und Wechselwirkungen sind auf der y-Achse nach ihrem Mittelwert sortiert. Die mit Doppelpunkt verbundenen Parameterwerte stehen für die entsprechenden Interaktionen. Die Parameterwerte P_L = `shortest`, P_T = `tree_none`, P_F = `none`, *representation* = `no` und *case* = `lc` bilden jeweils die Referenzniveaus, liegen also *per definitionem* bei 0. Die x-Achse zeigt $-ln(1/(1-P))$, mit der Performanz P.

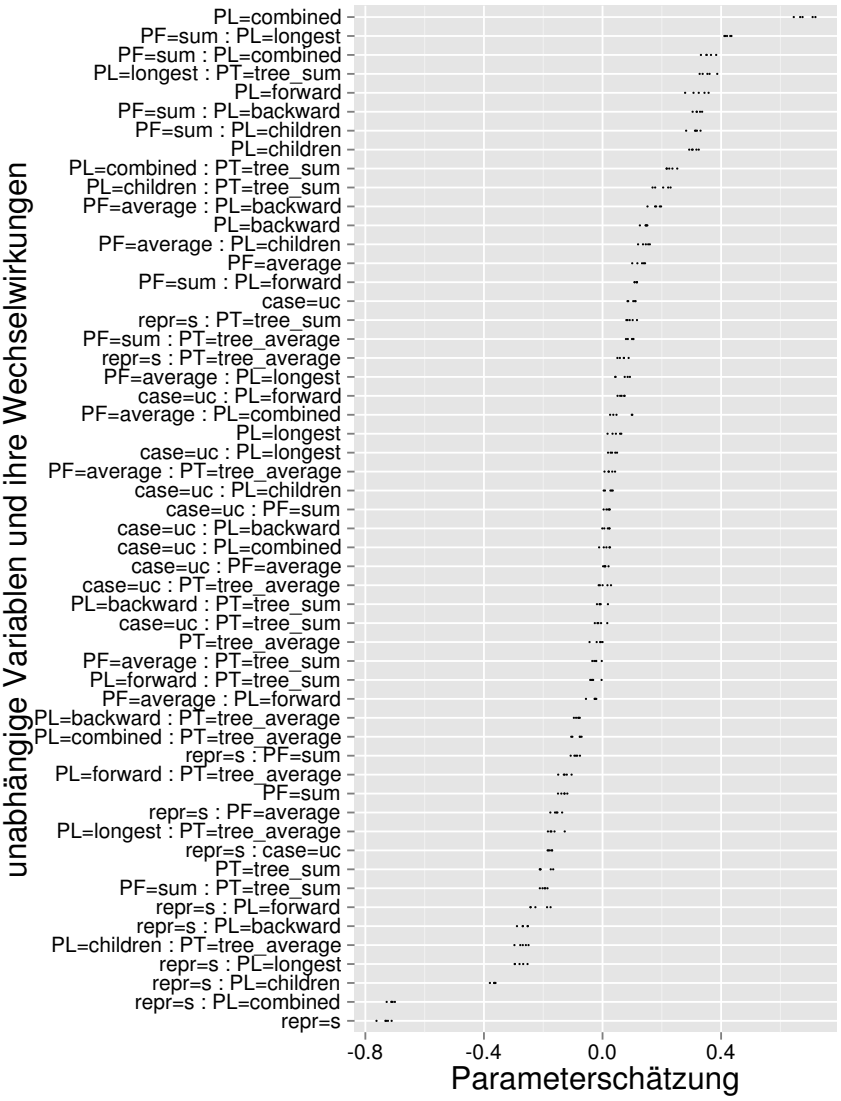


Abbildung 2.15: Die *englischen* Daten. Übersicht über die Parameterwerte, die sich für das im Text beschriebene gemischte generalisierte Modell ergeben. Positive/negative Werte auf der X-Achse korrespondieren mit einem positiven/negativen Einfluss der entsprechenden Parameterstellungen. Die Streuung in x-Richtung repräsentiert die Schwankung über die 5 Fitdurchläufe. Die Parameter und Wechselwirkungen sind auf der y-Achse nach ihrem Mittelwert sortiert. Die mit Doppelpunkt verbundenen Parameterwerte stehen für die entsprechenden Interaktionen. Die Parameterwerte P_L = `shortest`, P_T = `tree_none`, P_F = `none`, *representation* = `no` und *case* = `lc` bilden jeweils die Referenzniveaus, liegen also *per definitionem* bei 0. Die x-Achse zeigt $-ln(1/(1-P))$, mit der Performanz P.

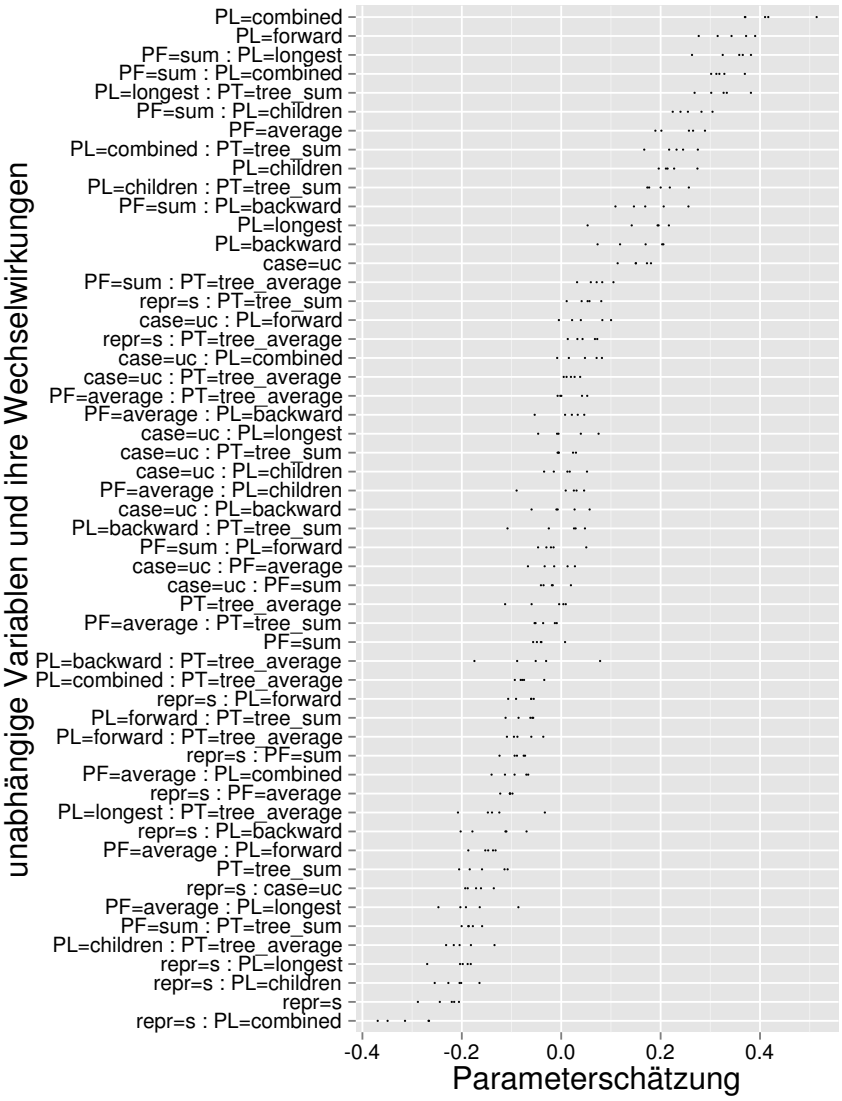


Abbildung 2.16: Die *türkischen* Daten. Übersicht über die Parameterwerte, die sich für das im Text beschriebene gemischte generalisierte Modell ergeben. Positive/negative Werte auf der X-Achse korrespondieren mit einem positiven/negativen Einfluss der entsprechenden Parameterstellungen. Die Streuung in x-Richtung repräsentiert die Schwankung über die 5 Fitdurchläufe. Die Parameter und Wechselwirkungen sind auf der y-Achse nach ihrem Mittelwert sortiert. Die mit Doppelpunkt verbundenen Parameterwerte stehen für die entsprechenden Interaktionen. Die Parameterwerte P_L = `shortest`, P_T = `tree_none`, P_F = `none`, *representation* = `no` und *case* = `lc` bilden jeweils die Referenzniveaus, liegen also *per definitionem* bei 0. Die x-Achse zeigt $-ln(1/(1-P))$, mit der Performanz P.

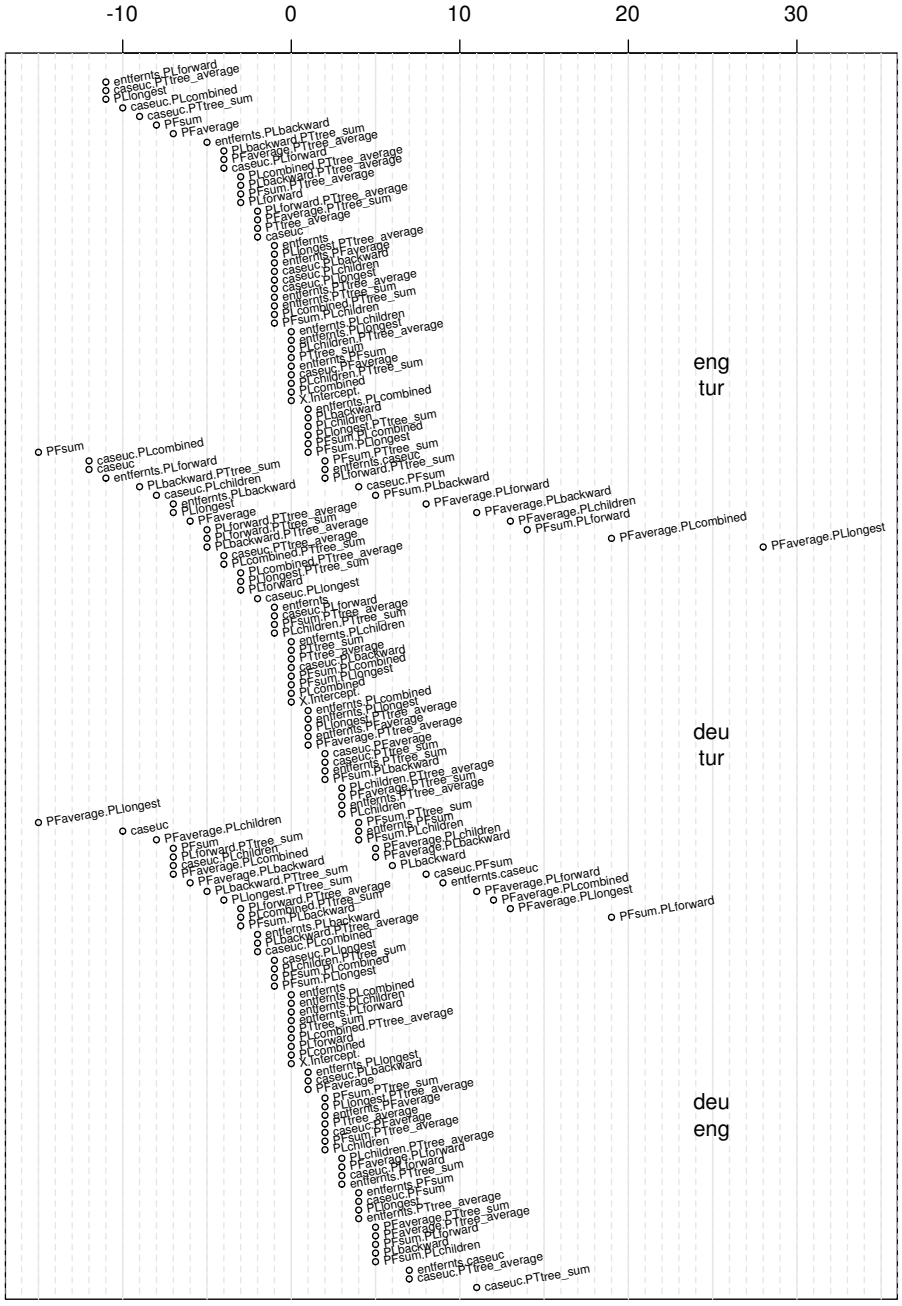


Abbildung 2.17: Platzwechsel aus dem Vergleich von Abbildung 2.14 bis 2.16. Ein Beispiel zur Bedeutung der Datenpunkte: Die Wechselwirkung von $P_F =$ `average` und $P_L =$ `longest` nimmt im deutschen Korpus (Abbildung 2.14) von oben gezählt die 36. Stelle ein. In den englischen Daten (Abbildung 2.15) steht sie an 21. Stelle. Entsprechend ist sie hier im obersten Teilbild, wo die deutschen mit den englischen Daten verglichen werden, bei $x = 21 - 36 = -15$ eingetragen.

Türkisch tendenziell sehr ähnlich verhalten, während das Deutsche abweichen sollte: Türkisch und Englisch folgen in Bezug auf Groß- und Kleinschreibung sehr ähnlichen Orthographieregeln, im Gegensatz zum Deutschen.

Die zweite Vermutung betrifft die absolute Richtung des Effektes für Englisch und Türkisch. Da die Großschreibung hier ein deutliches Signal für die Wortart (Eigenname) ist, sollte eine Nivellierung von Groß- und Kleinschreibung zu einem Informationsverlust und infolgedessen eher zu einer Performanzverminderung als einer -erhöhung führen. Die dritte Vorhersage betrifft die relative Lage von Deutsch im Vergleich zu Englisch und Türkisch: Hier gibt es neben dem negativen Effekt des Verlustes der Information über die Wortart den positiven Effekt durch die Verbesserung der Statistik für *minimale sprachliche Zeichen*, die durch groß und klein geschriebene Zeichenketten realisiert werden können. Es ist daher zu erwarten, dass die Nivellierung der Schreibweise hier im Vergleich zu den beiden anderen Sprachen positive Folgen haben sollte.

Die drei Vorhersagen treffen zu: Der Einfluss der beiden Ausprägungen von `case` in Englisch und Türkisch ist sehr ähnlich. Sowohl in Abbildung 2.15 als in Abbildung 2.16 findet sich *case* = `uc` deutlich im positiven Bereich. Das Deutsche weicht davon deutlich ab. In Abbildung 2.14 ist der Effekt von *case* = `uc` nicht von 0 unterscheidbar.

Es sind noch viele weitere Untersuchungen notwendig, bis die hier präsentierten Daten und die aus ihnen gezogenen Schlussfolgerungen als bestätigt oder gar gesichert gelten können, und vor allem bevor aus ihnen weitere Überlegungen in Bezug auf die Dynamik menschlicher Sprache bzw. von Lesen und Schreiben gezogen werden können. Zwei Untersuchungen bieten sich unmittelbar an. Zum einen sollten dänische und norwegische Texte aus der Zeit „deutscher“ Großschreibung (s. Fussnote 67) untersucht werden. Sie sollten sich verhalten wie der hier untersuchte deutsche Text. Ähnliches gilt für Texte aus Sprachen wie Französisch oder Italienisch, in denen *case* die Wirkung haben sollte, die hier für Englisch und Türkisch beobachtet wurde. Eine weitere mögliche Untersuchung würde in deutschen Texten lediglich die Großschreibung der Eigennamen nivellieren. Aus den bisherigen Ergebnissen lässt sich schließen, dass dies einen negativen Effekt haben sollte.

Nach der Behandlung von *representation* und *case* folgt nun die Diskussion der drei Parameter P_L, P_T und P_F und ihrer Wechselwirkungen.

Der Parameter P_L: Von der Beschreibung der Parameter in 2.5.2 ausgehend kann vermutet werden, dass der erste Parameter P_L, der festlegt, auf welche Art der *lokale Güteindex* vergeben wird, der einflussreichste sein dürfte: Er bestimmt die erste Bewertungsstufe, auf der alle weiteren aufbauen. Diese Einschätzung wird durch die Tatsache bestätigt, dass die Mittelwerte für die verschiedenen Werte von P_L etwa 2% schwanken, während P_T und P_F einen 2 bis 5 mal kleineren Einfluss auf die Performanz haben.

Ich beginne die Diskussion daher mit P_L. Die relative Stellung der möglichen Parameterwerte für P_L ist in allen Sprachen gleich: P_L = `shortest` schneidet jeweils am schlechtesten ab, dann folgen `longest`, `backward`, `children`, `forward` und – mit deutlichem Abstand – `combined`.

Die Spitzenstellung von `combined` kann man als Bestätigung interpretieren, dass die

Definition der *möglichen Segmente* über eine Kombination aus *forward* und *backward predictability change* eine sinnvolle Wahl war: Die Summe der Logarithmen beider Größen erlaubt die effektivste Bestimmung der Wortgrenzen. In dieses Bild passt auch, dass zwei Varianten, die jegliche Frequenzinformation ignorieren, `shortest` und `longest`, systematisch am schlechtesten abschneiden.

Es könnte auf eine interessante Asymmetrie in den Daten hindeuten, dass der *forward predictability change* alleine besser funktioniert als sein Spiegelbild, der *backward predictability change*. Im Rahmen der Analyse eines kleinen deutschen Goldstandard (2.6.3) wird sich im Zusammenhang mit dem bisher vernachlässigten Parameter P_4 ein interessanter zusätzlicher Aspekt in Bezug auf den Unterschied zwischen *forward* und *backward predictability change* ergeben.

Bemerkenswert ist die Größe des Abstandes zwischen `combined` und `forward` oder `backward`. Die Kombination beider Teilinformationen bedeutet eine erhebliche Verbesserung. Dies könnte einerseits daran liegen, dass sich *forward* und *backward predictability change* gegenseitig sehr effektiv ergänzen. Eine alternative Erklärung kommt aus den Berechnungsdetails des *lokalen Güteindex*: Im Falle von $P_L =$ `forward` bzw. `backward` wird jeweils der untransformierte *forward/backward predictability change* direkt verwendet, während im Falle von $P_L =$ `combined` die Summe der **Logarithmen** in den *Güteindex* eingeht. Falls dieser Unterschied in der Berechnungsweise für den Unterschied in der Performanz verantwortlich ist, sollte dieser nur in den Fällen existieren, in denen P_T oder P_F überhaupt eine Kombination der Werte für verschiedene Segmente implizieren. In $P_F =$ `none` bzw. $P_T =$ `tree_none` ist dies nicht der Fall (Vergleiche auch Abbildung 2.6, die das Verhältnis der drei Parameter visualisiert). Betrachtet man Abbildung 2.12, so fällt auf, dass genau für diese Einstellung der Unterschied zwischen `forward` und `combined` verschwindet. Dies erkennt man im oberen linken Teilbild. Im zentralen Teilbild dagegen, das der Stellung $P_F =$ `sum` und $P_T =$ `tree_sum` entspricht, ist der Unterschied maximal. Dies ist nun aber genau der Fall, in dem die Bewertungen der einzelnen Segmentkandidaten als Summanden in die Gesamtbewertung einfließen.

Dies stützt die Vermutung, dass die Ursache der starken Überlegenheit von $P_L =$ `combined` gegenüber `forward` und `backward` tatsächlich auf die Verwendung des Logarithmus zurückzuführen ist. Diese Beobachtung korrespondiert mit Beobachtungen, die wir im zweiten Teil der Arbeit im Bereich der Stilometrie machen werden (Kapitel 3). Dieser empirische Befund einer allgemeinen Überlegenheit des Logarithmus ist einer der möglichen Berührungspunkte zwischen den hier gemessen Performanzunterschieden eines anwendungsorientierten Algorithmus und dem dynamischen System der Sprache wie es in der Einleitung (Kapitel 1) diskutiert wurde.

Interessant, aber auch gut nachvollziehbar ist die Tatsache, dass `children`, diese sehr einfache Berechnungsmethode des *lokalen Güteindex* so gut funktioniert: Hier entspricht der Güteindex direkt der Zahl der Blätter unterhalb eines Segments. D.h., die reich verzweigten Bäume entsprechen recht gut den linguistisch sinnvollen, sofern man die Zahl der erkannten Leerzeichen als Index hierfür akzeptieren möchte. Kurz: Wo viel gefunden wird, ist man auf dem richtigen Weg. `children` hebt sich von den übrigen Werten für P_L ab, da hier gerade eben nicht nur Informationen eines Segmentes für sich

betrachtet werden, sondern die Zahl seiner Kinder (und Kindeskinder).[71]

Der Parameter P_T: Von P_L gehe ich nun über zum Einfluss von P_T, der festlegt, wie die Einzelbewertungen der Segmente zur Bewertung ganzer Segmentbäume kombiniert werden. Auch P_T zeigt ein in allen drei Sprachen identisches Muster. Am besten ist es, für die Bewertung eines Segmentes nur dieses selbst zu beachten, und nicht die Bewertungen seiner Kindsegmente: $P_T =$ `none` schneidet besser ab als `tree_average` oder – mehr noch – als `tree_sum`.

Dies ist nur ein scheinbarer Widerspruch zum optimalen Parametersatz wie in Abbildung 2.10 berichtet. Dort gilt $P_T =$ `tree_sum`. Die übrigen beiden Parameter sind hier allerdings $P_L =$ `combined` und $P_F =$ `sum`. In Abbildung 2.14 ist zum Beispiel zu erkennen, dass zwischen $P_T =$ `tree_sum` und $P_L =$ `combined` eine recht starke positive Interaktion besteht. Dies ist ein Beispiel für das auf Seite 88 beschriebene Problem, die Haupteffekte ohne Berücksichtigung vorhandener Interaktionen zu interpretieren. Die Überlegenheit von $P_T =$ `none` ist real, sie gilt aber nur gemittelt über die übrigen Parameter. Sie deutet, wie ja auch schon die Überlegenheit von `longest` über `shortest` und auch das gute Abschneiden von `children` darauf hin, dass die Segmente höher im Baum, bzw. die längeren Segmente, oder die Segmente mit den meisten Kindsegmenten am erfolgreichsten identifiziert werden können.

Der Parameter P_F: Der Parameter P_F wiederum hat in der Horizontalen die Bedeutung, die P_T in der Vertikalen hat: Während P_T festlegt, wie die Kindsegmente eine *Segmentes* in dessen Bewertung eingehen, so bestimmt P_F die Bewertung der Folgesegmente. P_F zeigt im Gegensatz zu P_L und P_T ein von Sprache zu Sprache variables Muster. Zwar gilt, dass $P_F =$ `average` immer deutlich von Vorteil gegenüber `none` ist. Diese Wahl ist wiederum besser als `sum`, wobei dieser Unterschied im Türkischen kleiner ist.

Für die Beobachtung, dass $P_F =$ `average` gegenüber den anderen Möglichkeiten überlegen ist, lässt sich *a posteriori* auch eine Erklärung finden: Hier wird ein Durchschnitt über alle folgenden Segmente zur Bewertung der Gesamtsegmentierung herangezogen. Dadurch werden nicht besonders viele Folgesegmente als gut angesehen, wie es bei schlichter Summierung oft[72] der Fall ist (`sum`). Statt dessen wird eine Folgereihe möglichst gut bewerteter Segmente herangezogen. Da dies nach den bisherigen Erkenntnissen eher die langen sind, ist `average` gegenüber `sum` die bessere Wahl.

Zusammengefasst: Bei P_T wirkt sich Durchschnittsbildung tendenziell ungünstig aus, da die längeren Segmente auf höherer Ebene oft die besseren sind. Bei P_F ist Mittelung

[71] Eine alternative Betrachtungsweise wäre es, eine Einstellung für P_L zu definieren, in der der Güteindex I_L eine Konstante ist ($I_L(s) = 1$ für alle s). Dann wäre $P_T =$ `sum` gleichbedeutend mit der jetzigen Kombination $P_L =$ `children` und $P_T =$ `tree_none`

[72] Hier erweist sich die Tatsache als suboptimal, dass I_L je nach Wahl von P_L mal positiv und mal negativ ist. Da der Güteindex aber immer minimiert wird, führt dies notwendigerweise zu Wechselwirkungen mit P_F. Für den wichtigen Fall $P_L =$ *combined* sind die summierten Logarithmen immer negativ, da $D < 1$ eine Vorraussetzung für den Status als *mögliches* Segment ist.

dagegen eher positiv, weil die Segmente bevorzugt werden, die sich besonders überzeugend fortsetzen lassen.

Auch wenn die relative Reihenfolge der möglichen Werte für P_F über die Sprachen ziemlich stabil ist, so ist dieser Parameter dennoch einer wesentlich weiteren Streuung unterworfen als P_T.

Abbildung 2.17 verdeutlicht dieses Phänomen. Diese Graphik gibt einen Überblick, welche Parameter in ihrem relativen Einfluss von Sprache zu Sprache schwanken: Hat ein Parameter in zwei Sprachen den maximalen positiven Einfluss, so bekommt er in beiden Fällen den Rangplatz 1 zugewiesen. Entsprechend ist die Differenz der Rangplätze für diesen Parameter für dieses Sprachpaar 0. Für alle drei Paarungen sind die entsprechenden Rangplatzunterschiede dargestellt.

Ich beginne die Diskussion mit einer Plausibilitätsprüfung. Entsprechend der Verwandtschaftsverhältnisse und der typologischen Charakteristika ist zu erwarten, dass sich die beiden westgermanischen Sprachen Englisch und Deutsch untereinander ähnlicher verhalten, als jeweils im Vergleich mit dem Türkischen.

Die Abbildung reproduziert diese Erwartung. Während zwischen Deutsch und Englisch kein Haupteffekt und keine Interaktion um mehr als 15 Plätze schwankt, so ist dieser Schwankungsbereich für die Vergleiche Türkisch-Deutsch und Türkisch-Englisch mit 19 bzw 28 Plätzen deutlich höher.

Nun komme ich wieder auf den von Sprache zu Sprache schwankenden Einfluss des Parameters P_F zurück. Diese Variabilität von P_F kann ebenfalls aus Abbildung 2.17 abgelesen werden. Die Parameterwerte P_F = `sum` und `average` finden sich fast ausschließlich im äußeren linken Bereich, sie wechseln also intensiv die Plätze verglichen mit den anderen Parametern. Nur im Englisch-Deutschen Vergleich verändert wenigstens P_F = `average` seinen Platz kaum.

Die entsprechenden Werte für P_T=`sum` und `average` dagegen befinden sich ausschließlich im innersten Bereich, haben also in allen drei Modellen dieselbe Stellung unter den untersuchten Parametern.

Interaktionen: Sehr stark sprachabhängig ist die Wechselwirkung zwischen P_F und P_L. Diese beherrscht für Deutsch-Englisch den linken äußeren Bereich und für die beiden anderen Vergleiche den rechten äußeren.

Die Sprachabhängigkeit der übrigen Interaktionen zwischen P_L und P_T und zwischen P_T und P_F ist demgegenüber kleiner. Diese Kombinationen sind jeweils weiter von den äußersten Bereichen der Graphiken in Abbildung 2.17 entfernt.

Diese deutliche Sprachabhängigkeit von P_F deutet darauf hin, dass an dieser Stelle nach linguistisch bedeutsamen Zusammenhängen zwischen den recht unterschiedlichen morphologischen Strukturen der Sprachen, den hier ausgenutzten Frequenzinformationen und den im Algorithmus variierten Parametern gesucht werden könnte.

Unbesprochen bleiben in dieser Diskussion die Wechselwirkungen zwischen *case* und *representation* auf der einen Seite und den übrigen Parametern auf der anderen Seite. Diese schien mir vor allem linguistisch nicht besonders interessant zu sein.

Ich beende diesen Abschnitt mit einer kurzen Zusammenfassung der bisherigen em-

pirischen Ergebnisse:

Der Algorithmus ist in der Lage bis über 95% der Leerzeichen im Text als Morphemgrenzen zu identifizieren.

Die Leerzeichen aus dem Text zu entfernen, behindert die Segmentierung erheblich.

Die Groß- und Kleinschreibung zu normalisieren hat sprachabhängige Folgen, die ohne weitere Untersuchungen nur schwer zu deuten sind. Mit Blick auf mögliche Anwendungen bringt die Normalisierung aber keinen sehr weitgehenden Vorteil.

P_L ist ein einflussreicher Parameter, der sich im wesentlichen sprachunabhängig verhält. Die Bewertungsmethode, die die Frequenzinformation am vollständigsten ausnutzt (`combined`) hat die beste Performanz.

Mindestens so stabil wie P_L ist P_T. Sein Verhalten und sein Zusammenspiel mit P_L lässt sich *a posteriori* in wichtigen Punkten einleuchtend erklären.

Stark sprachabhängig ist dagegen P_F. Auch die Interaktionen dieses Parameters mit P_L schwanken stark von Sprache zu Sprache.

Beide Beobachtungen, die Stabilität von P_L und P_T, als auch die Sprachabhängigkeit von P_F eröffnen neue Fragen und suggerieren weitere Untersuchungen. Gelten beide Eigenschaften auch für weitere, bestenfalls typologisch stark abweichende Sprachen? Wenn es Unterschiede geben sollte, mit welchen bekannten Eigenschaften der Sprachen korrespondieren sie?

Die Ergebnisse dieses Abschnitts sind also zweierlei: Zum einen ergibt sich eine Fülle von neuen empirischen Fakten, die sich teilweise klar auf bekannte linguistische Tatsachen zurückführen lassen. Dies ist zumindest ein Konsistenztest. Zum anderen gibt es eine Gruppe weiterer Beobachtungen, für die sich plausible linguistische Erklärungen finden lassen, ohne dass diese sich aufgrund der bisherigen Datengrundlage strikt beweisen lassen. Für andere Befunde wiederum muss ich eine Deutung schuldig bleiben. Die Daten werden dadurch meines Erachtens nicht uninteressant, sondern wird im Gegenteil ihr Reichtum dadurch unterstrichen.

2.6.3 Evaluation eines kleinen Goldstandard

In den vorangegangenen Abschnitten werden nur Segmentgrenzen betrachtet, die mit Leerzeichen zusammenfallen. Die Motivation für diese Einschränkung ist, dass sich so eine große Menge gesicherter Positivbeispiele gewinnen lässt. Zwei Nachteile eines solchen Vorgehens liegen jedoch auf der Hand: Erstens kann man auf diesem Weg keine Informationen über die Qualität des Algorithmus auf den weniger eindeutigen Grenzen im Wortinneren gewinnen. Es ist möglich, dass diese das Bild völlig verändern. Auf der anderen Seite existiert so nur Zugriff auf die Vollständigkeit, mit der vorhandene Positivbeispiele gefunden werden, den *Recall*, s. Definition 26, bzw. seine Verwandte, die *Performanz*, s. Definition 28. Die *Precision* (Definition 27), die Qualität der vorgeschlagenen Segmente, bleibt so unbekannt.

Um diese Probleme zu umgehen ist ein Goldstandard vonnöten, also eine ausreichende Menge Text, in der sämtliche Segmentgrenzen bekannt sind. Leider ergeben sich weitere Probleme. Die unvermeidliche Theorieabhängigkeit eines morphologischen Goldstandards wurde bereits auf Seite 73 angesprochen. Ein eher technisches Hindernis ist der

große Aufwand, den es erfordert, größere Mengen Text sorgfältig zu annotieren.

Aus diesem Grund habe ich mich auf einen sehr kleinen Goldstandard beschränkt. Er umfasst lediglich 20 Sätze des deutschen Testkorpus. Erstaunlicherweise wird sich zeigen, dass dies ausreicht, um recht präzise Aussagen zu treffen.

Die Theorieabhängigkeit wurde quantitativ berücksichtigt. Drei ausgebildete Linguisten[73] segmentierten den Text unabhängig voneinander. Theoretische Vorgaben wurden auch auf Nachfrage verweigert. So ist die Breite der theoriebasierten Variation zumindest abschätzbar.

Insgesamt kamen so 840 Segmentgrenzen zusammen, oder durchschnittlich 42 pro Satz. Von diesen existierten 714 oder 85% übereinstimmend in allen drei Segmentierungen. In 78 Fällen (9%) waren sich immerhin 2 der Experten einig, 48 Segmentgrenzen (6%) wurden nur von einem der Befragten gesetzt.[74]

Wie erwartet stimmen alle Annotatoren in allen Fällen darin überein, dort Segmentgrenzen zu setzen, wo nach der deutschen Orthographie Leerzeichen geschrieben werden. Auch in Bezug auf transparenten und produktiven Prozessen entstammende Bildungen wie

```
unter der Herr schaft des Sozial ist en ge setz es
```

herrscht Einigkeit.

Unterschiede gibt es unter anderem bei sehr etablierten und nicht mehr so transparenten Bildungen. Ein Beispiel ist „Erinnerung“, das als

```
Er inner ung
Erinner  ung
```

segmentiert wird. Die erste Segmentierung ist aus diachroner Hinsicht korrekt, da das Wort über die Bildungsregel `er` + $ADJ \rightarrow VERB$ entstanden ist. Diachron betrachtet allerdings ist diese Ableitung aber allenfalls noch marginal produktiv. Das heißt, je nach Blickpunkt ist entweder die eine oder die andere Zerlegung korrekt. Ein ähnliches Beispiel ist „Urteil“, das als

```
Ur teil
Urteil
```

segmentiert wird und „dazu“, das in den folgenden Varianten auftritt:

```
da zu
dazu
```

Qualitativ unterschiedlich verhält es sich zum Beispiel mit Suppletionen wie „besser“, das in die Varianten

[73] Dank an Hagen Hirschmann, Marc Reznicek und Amir Zeldes.

[74] Es ist im Lichte dieser Zahlen nicht ganz uninteressant, dass sich die drei befragten Linguisten kennen, und seit Jahren in derselben Arbeitsgruppe arbeiteten. Das lässt vermuten, dass mit steigendem räumlichem und fachlichem Abstand der befragten Experten auch die Diskrepanz der Urteile eher noch weiter zunehmen wird.

```
besser
bes ser
bess er
```

segmentiert wird. Hier ist es sehr unabhängig vom Beschreibungsrahmen sehr schwierig, oder gar unmöglich, eine gültige Lösung zu finden. Auch die genaue Trennung komplexer Endungen wird teilweise von allen drei Annotatoren unterschiedlich gesehen: „einem Zeugen“ wird als

```
einem Zeug en
einem Zeug e n
einem Zeuge n
```

segmentiert. Ähnliche Unterschiede gibt es bei Portmanteaumorphen wie „hatte“, das als

```
hat t e
hatt e
hat te
```

getrennt wird. Das zweite `t` hat hier doppelte Funktion. Zum Einen dient es der Verdeutlichung der kurzen Quantität des `a`, zum anderen ist es Teil der Präteritumsendung.

Um die Varianz dieser Daten zu quantifizieren, definiere ich die *Sicherheit* einer Segmentgrenze:

Definition 29 (Sicherheit) *Die Sicherheit C_i einer Segmentgrenze i ist der Anteil an Experten, die diese Segmentgrenze gesetzt haben:*

$$C_i = \frac{D_i}{E_i}$$

mit der Expertenzahl E_i und der Zahl der positiven Entscheidungen D_i.

In unserem Fall ist E_i eine Konstante ($E_i = 3$). Es existieren die *Sicherheit*swerte 0, $\frac{1}{3}$, $\frac{2}{3}$ und 1.

Eine solche Beschreibung ist nicht ohne Probleme. Es ist möglich und bis zu einem gewissen Grad auch anzunehmen, dass jeder der befragten Experten in sich über eine konsistente Grammatik verfügt. Definition 29 mittelt über diese verschiedenen, aber in sich geschlossenen, Grammatiken. Das Ergebnis ist eine Mischgrammatik. An dieser Mischgrammatik bzw. dem daraus entstehenden Goldstandard wird der Algorithmus gemessen. Dabei wäre es vielleicht eine noch interessantere Frage, welcher der Expertengrammatiken der Algorithmus am nächsten kommen kann. Für eine solche Untersuchung wäre aber ein Vielfaches an Daten notwendig. Dies bezieht sich nicht nur auf die Menge annotierten Textes, die nötig wäre, die Konsistenz der einzelnen Annotatoren zu überprüfen, sondern auch auf die Zahl der Annotatoren. Drei Experten sind bei weitem zu wenig, um tragfähige Aussagen über die Varianz zwischen Personen und Grammatiksystemen machen zu können. So bleibt nichts, als die in den Daten sichtbare Theorieabhängigkeit auf die relative Sicherheit einzelner Entscheidungen zu reduzieren. Entscheidungen, die von allen Annotatoren getroffen werden, können als vertrauenswürdig und

theorieunabhängig gelten. Wenn Segmentgrenzen dagegen nur von einem oder zwei Annotatoren gesetzt wurden, können sie als theorieabhängiger angenommen werden. Diesen eine geringere *Sicherheit* zuzusprechen, bedeutet, dass der Algorithmus vor allem an den unstrittigen Segmentgrenzen gemessen wird, die theorieabhängigeren Segmente aber nicht völlig ignoriert. Die Varianz beziehungsweise die Unbestimmtheit der Segmentgrenzen auf diese Art und Weise mit einzubeziehen scheint insgesamt durchaus sinnvoll. Es folgen daraus allerdings neue Schwierigkeiten. So setzen die Definitionen von *Recall* und *Precision* voraus, dass es nur eindeutige Positivbeispiele gibt.

Ich modifiziere sie dementsprechend, um Maße zu gewinnen, die die Sicherheit bzw Unsicherheit der verschiedenen Instanzen im Goldstandard berücksichtigt:

Definition 30 (weighted Recall)

$$r_w = \frac{\sum_{i=1}^{n} C_i}{m}$$

Der Index i läuft hier über alle Segmentgrenzen i, die vom zu evaluierenden System gesetzt wurden. m ist die Zahl der Segmentgrenzen im Goldstandard.

Für einen herkömmlichen Goldstandard, in dem es nur Positiv- und Negativbeispiele gibt ($C_i \in \{0, 1\}$), fällt diese Definition mit dem üblichen *Recall* zusammen.

Entsprechend ist die *weighted Precision* definiert:

Definition 31 (weighted Precision)

$$p_w = \frac{\sum_{i=1}^{n} C_i}{p}$$

Der Index i läuft wieder über alle Segmentgrenzen i, die vom zu evaluierenden System gesetzt wurden. p ist die Zahl der vom System vorgeschlagenen Grenzen.

Auch diese Definition geht für einen herkömmlichen Goldstandard in die übliche *Precision* über.

Häufig wird statt der getrennten Evaluationsmaße *Recall* und *Precision* ihr harmonisches Mittel betrachtet. Entsprechend definiere auch ich:

Definition 32 (weighted f-measure)

$$f_w = \frac{2 p_w r_w}{p_w + r_w}$$

Man kann es als Nachteil dieser abgewandelten drei Evaluationsmaße sehen, dass ihr Wertebereich im Allgemeinen unterhalb von 1 endet. So ist er für einen Goldstandard, der nur aus Segmentgrenzen der *Sicherheit* 1/2 besteht, auch maximal 1/2. Diese Eigenschaft birgt aber auch eine gewisse Vernunft in sich. So kann man argumentieren, dass

es für einen Goldstandard, der graduelle Wertungen zwischen 0 und 1 enthält, keine perfekte automatisierte Lösung geben kann, falls das System nicht ebenfalls zu graduellen Entscheidungen fähig ist. Es soll an dieser Stelle noch einmal betont werden, dass die Graduiertheit nicht den Urteilen der einzelnen Annotatoren entstammt. Diese waren gezwungen, sich bestimmte Segmentgrenzen zu entscheiden. Zwischenwerte entstehen aus der Varianz *zwischen* den Annotatoren.

Für den hier vorliegenden Goldstandard liegt der maximale *weighted Recall* bei $\left(\frac{714 \cdot 1 + 78 \cdot \frac{2}{3} + 48 \cdot \frac{1}{3}}{840} = 0.93\right)$.

Für die maximale *weighted Precision* ist die Frage nach dem Maximalwert nicht so eindeutig, da nicht unmittelbar klar ist, was im Fall eines teilweise unsicheren Goldstandards als „perfekte Lösung“ zählen soll. Betrachtet man die Menge der Segmentgrenzen der *Sicherheit* 1 als die optimale Lösung, so liegt auch die maximale *weighted Precision* für unseren Goldstandard bei 1. Akzeptiert man aber die Menge aller Grenzen der *Sicherheit* > 0 als optimale Lösung sinkt die maximale *weighted Precision* auch auf 0.93 ab.

Entsprechend liegt das maximale *weighted f* bei $\left(\frac{2 \cdot 0.93}{0.93 + 1} = 0.96\right)$ oder ebenfalls bei nur 0.93.

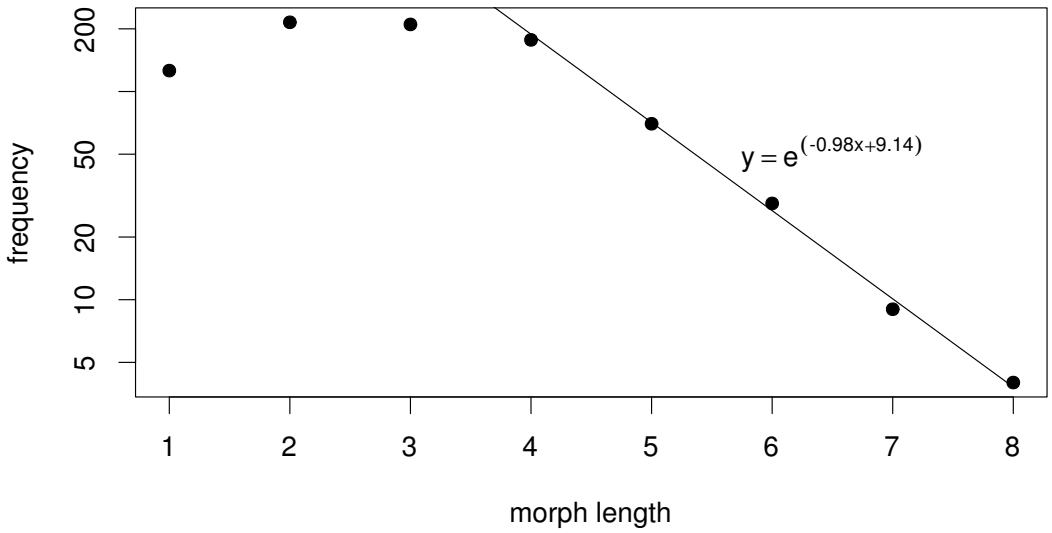


Abbildung 2.18: Längenverteilung der Goldstandard-Morphe. Die durchgezogene Linie repräsentiert eine Regressionsgerade durch die rechten vier Punkte. Die Steigung der Geraden ist mit -1.0 verträglich. Der exponentielle Schwanz der Verteilung legt eine Poissonverteilung nahe. Um durch eine Poissonverteilung modellierbar zu sein, müsste die Varianz aber gleich dem Mittelwert sein. Der Mittelwert der Verteilung ist nicht von 3 zu unterscheiden, die Varianz beträgt jedoch nur ziemlich genau 2.

Abbildung 2.18 ist inhaltlich ein Exkurs. Hier ist die Längenverteilung der von den be-

fragten Experten annotierten Segmenten aufgetragen. Die y-Achse ist logarithmisch.[75] Auffällig ist der recht exakt exponentielle Abfall für größere Längen. Der Koeffizient ist wohl zufälligerweise gut verträglich mit einem Wert von genau -1. Dieser exponentielle Schwanz der Verteilung ist ein Einwand gegen die von Altmann und Best vorgetragene Hypothese, dass Wort- und Morphlängen sich durch die Hyperpoissonverteilung beschreiben lassen (s. z.B. Best, 2001). In einem ähnlichen Modell schlagen Creutz (2003) bzw. Creutz und Lagus (2002) direkt die Poissonverteilung zur Modellierung von Wort- und Morphlängen vor. Beide Verteilungen fallen für große Wortlängen schneller ab als die Exponentialverteilung. Für die Poissonverteilung wird das unmittelbar einsehbar, wenn man sich vor Augen führt, dass sie für ausreichen große Erwartungswerte durch die Normalverteilung genähert werden kann, die wie e^{-x^2} abfällt. Der Irrtum rührt möglicherweise daher, dass die Diskrepanz im rechten Teil der Verteilung nur in logarithmischer Darstellung sichtbar wird. Der von den zitierten Autoren verwendete χ^2-Test kann derartige Abweichungen nicht aufdecken.

Wenden wir uns nun der quantitativen Analyse der drei *weighted* Evaluationsmaßen in Abhängigkeit der Parameter zu. Dies werden zuerst einmal wie bisher nur *representation*, *case*, $P_{L,F,T}$ sein. Erst in einem letzten Schritt wird dann auch P_4 mit einbezogen.

Es ist eine sinnvolle Annahme, dass manche Sätze für das System durchgängig schwerer zu segmentieren sind als andere. Zu dieser Varianz auf Satzebene hat man Zugang, wenn die *weighted*-Werte nicht jeweils für den gesamten Testtext berechnet werden, sondern für die einzelnen Sätze.

Damit gibt es für jedes der drei Evaluationsmaße und für jede Parameterstellung 20 Werte, für jeden Satz einen. Insgesamt sind es 3×56160 Datenpunkte.[76]

Wir stehen nun vor einer sehr ähnlichen Fragestellung wie in Abschnitt 2.6.2, wo es um die Modellierung der *Performanz* ging. Entsprechend wird sich eine vergleichbare Antwort anbieten.

Die drei Voraussetzungen an ein Modell mit normalverteilter Fehlervarianz wurden dort bereits erwähnt: Unabhängigkeit der Datenpunkte, Normalverteilung der Residuen und Unabhängigkeit der Fehlervarianz von den Variablen.

Für die Analyse der Leerzeichen verbietet sich die Annahme normalverteilter Residuen von vornherein, sowohl aufgrund visueller Inspektion, als auch durch theoretische Überlegungen. Hier ist die Sachlage nicht so klar: Die zugrunde liegende Verteilung dort war bestenfalls durch eine Binomialverteilung zu beschreiben und diese aufgrund ihrer spezifischen Parameter nicht durch eine Normalverteilung zu nähern.

Wie ist die Situation jetzt? Könnten die Voraussetzungen besser erfüllt sein? Ist es zum Beispiel vorstellbar, dass die *weighted* f-Werte für einen bestimmten Parametersatz um einen bestimmten Erwartungswert normalverteilt sind? Immerhin ist es durchaus möglich, dass die Verteilung einen größeren Abstand von 1 aufweist als die Verteilung der *Performanz*, da wir nun nicht nur die Wort(form)grenzen, sondern auch Segment-

[75] Der Kurve fehlen die bei Wortlängenverteilungen üblichen Unregelmäßigkeiten für kurze Längen, die gewöhnlich durch den übergroßen Einfluss einiger weniger (kurzer) Wörter hervorgerufen wird.

[76] 2 Werte für *representation*, 2 Werte für *case*, 6 Werte für P_L, 3 Werte für P_F, 3 Werte für P_T, 13 für P_4 und 20 Goldstandardsätze.

grenzen innerhalb von Wörtern betrachten. Durch die Mischung von Segmentgrenzen verschiedener *Sicherheit* ist die Vorstellung einer Binomialverteilung sowieso nicht mehr angemessen.

Es gilt also genauer hinzuschauen. Wären alle genannten Voraussetzungen erfüllt, wäre es möglich, den mächtigen mathematischen Apparat der *mixed linear models* auf die Daten anzuwenden. Die 5 (oder 6) Parameter, denen das Hauptinteresse gilt, würden als *feste Effekte* (fixed effects) modelliert. Dies heißt nichts anderes, als dass ihr Einfluss auf *weighted f* als systematisch und reproduzierbar angenommen wird. Darüber hinaus nehmen wir wieder an, dass es leicht und schwerer zu segmentierende Sätze gibt. Diese intrinsische Segmentierbarkeit wird als zwischen den Sätzen normalverteilt (als *zufälliger Effekt*, *random effect*) angenommen. Die Varianz dieser Normalverteilung geht als Parameter in das Modell ein. Derartige Modelle heißen *mixed*, da sie beide Klassen von Effekten zugleich beinhalten. Ich bevorzuge sie hier gegenüber der alternativen Möglichkeit einer multifaktoriellen Anova mit Messwiederholung wegen ihrer ungleich höheren Flexibilität.

Es gibt nur einen Weg, zu überprüfen, ob die Voraussetzungen für einen solchen Ansatz tatsächlich gegeben sind: Das Modell ist konkret durchzurechnen, damit anschließend die Verteilung der verbleibenden *Residuen* untersucht werden kann.

Eine erste Analyse gibt Abbildung 2.19. Aus den Graphiken kann geschlossen werden, dass die Residuen in der Tat in ausreichendem Maß normalverteilt sind. Die Güte der Übereinstimmung ist insofern erstaunlich, als die möglichen Werte für *weighted Precision, Recall* und *f* nach wie vor jeweils notwendigerweise zwischen 0 und 1 liegen müssen. Es ist aber offensichtlich tatsächlich so, dass die Performanz weit genug von 1 entfernt ist, als dass Randeffekte sichtbar würden.

Abbildung 2.20 zeigt einen anderen Aspekt der Residuenverteilung, der dann die Grenzen des verwendeten Ansatzes deutlich macht. Hier sind die Residuen über den vom Modell vorhergesagten Werten aufgetragen. Das auffällige Streifenmuster kommt daher, dass sich häufig für ein und denselben Satz unter verschiedenen Parameterstellungen identische Zerlegungen ergeben, zumindest auf der untersten Ebene, die hier ausgewertet wird. Dies ist nun leider doch eine klare Abweichung von der Unabhängigkeitsannahme. Das heißt, obwohl die Unabhängigkeit von Satz zu Satz wohl als gegeben angenommen werden kann, neigt *f* dazu, bei festen Werten gewissermaßen einzurasten.

Ein positiver Aspekt der Residuenverteilung in Abbildung 2.20 ist allerdings, dass die Varianz über einen breiten Bereich gefitteter Werte einigermaßen konstant ist.

Wir sind insgesamt in einer glücklicheren Lage als in 2.6.2 wo die Verwendung genau derartiger Modelle von vornherein verneint werden musste (s. Seite 86). Das verwendete Modell bleibt allerdings eine Näherung, wie jedes Modell. Es wird zu zeigen sein, dass diese Näherung konsistente und deutbare Ergebnisse liefert.

Ein weiterer Unterschied zur Leerzeichenanalyse des letzten Abschnitts liegt darin, dass es jetzt Argumente für die Verwendung herkömmlicher *p*-Werte gibt. **Möglich** ist die Verwendung der berechneten *p*-Werte wiederum auf Grundlage der Gültigkeit der Annahme mit gleichmäßiger Varianz normalverteilter Residuen. **Hilfreich** sind sie hier, da nun nicht mehr potentiell unendlich viele Daten zur Verfügung stehen, sondern lediglich 20 Sätze. Daher bedarf es eines Kriteriums um zu entscheiden, welchem Effekt

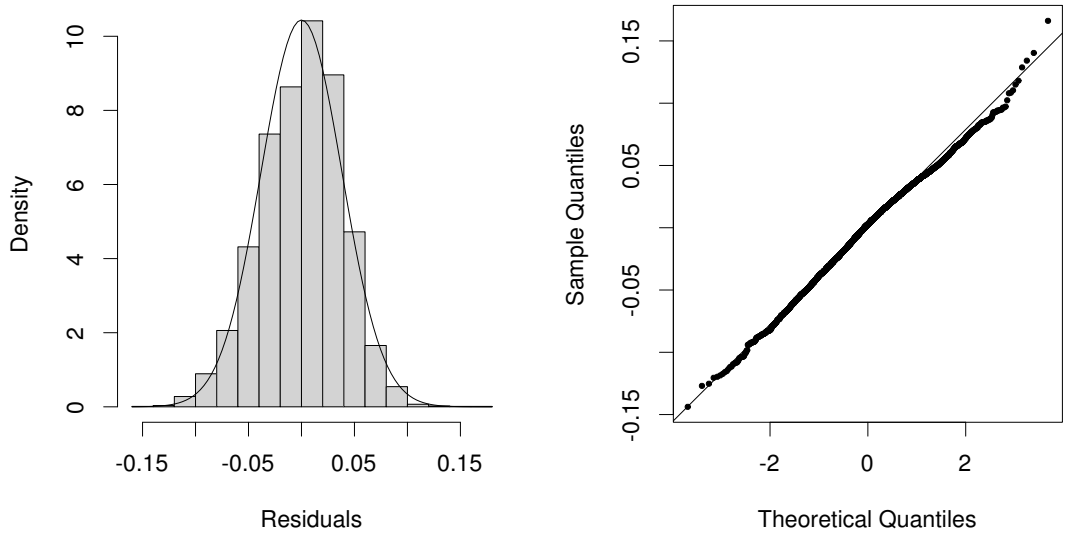


Abbildung 2.19: Verteilung der Residuen im optimalen *linear mixed* Modell. Linkes Teilbild: Das Histogramm der Residuen. Mit eingetragen ist eine Normalverteilung mit identischem Mittelwert und identischer Standardabweichung. Von einer leichten Verschiebung des Maximums nach rechts abgesehen ist die Übereinstimmung sehr gut. Gleiches kann aus dem rechts abgebildeten QQ-plot abgelesen werden. Eine Deckung von durchgezogener Linie und Residuen würde eine perfekte Übereinstimmung mit der Normalverteilung bedeuten.

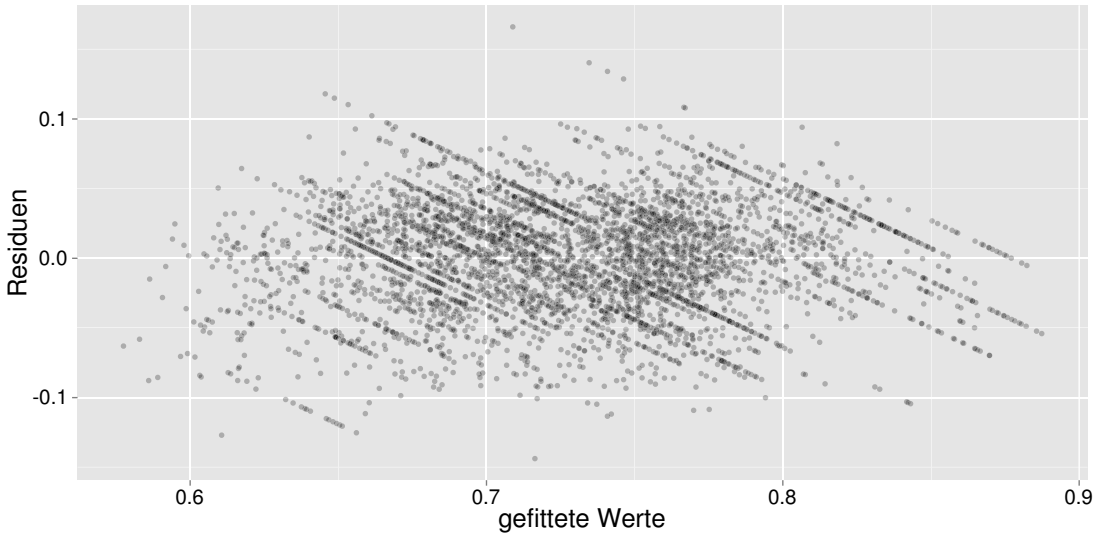


Abbildung 2.20: Residuenplot für das optimale Modell für *weighted* f.

bereits auf dieser Datengrundlage getraut werden kann und welcher nicht nachweisbar ist.

Ich beginne mit einem Vergleich mit den im vorigen Abschnitt 2.6.2 dargestellten Ergebnissen. Dort wurden ausschließlich Leerzeichen zur Auswertung herangezogen. Das definierte Evaluationsmaß der *Performanz* (Definition 28) ist, wie im Anschluss an diese Definition bereits ausgeführt, identisch mit dem *Recall* für diese eingeschränkte Menge an Morphemgrenzen. Daher sollten die Ergebnisse der damaligen Analyse mit einem Modell auf Grundlage des *weighted Recall* auf dem nun eingeführten Goldstandard zumindest qualitativ korrelieren.

Dies ist eine recht zuverlässige Konsistenzprüfung, da es beachtliche Unterschiede zwischen den beiden Untersuchungen gibt:

- Die Größenordnung der Datensätze unterscheidet sich um den Faktor 10.
- Die Qualität der Daten ist eine andere, da nun alle Segmentgrenzen einbezogen werden, nicht nur die Leerzeichen.
- Die Messgrößen sind zwar ähnlich, aber doch unterschiedlich: Bisher wurde die *Performanz* untersucht, nun ist es der *weighted Recall*.
- Das zugrundegelegte Modell ist unterschiedlich. Zur Analyse der Leerzeichen wurde eine binomiale Fehlerverteilung angenommen, nun legen wir eine normalverteilte Residuenverteilung zugrunde. Auch die verwendete Software ist eine andere. Für die Analyse der Leerzeichen wurde das `R`-Paket `lme4` (Bates et al., 2011) verwendet, der nun vorgestellten Untersuchung liegt das Paket `nlme` (Pinheiro et al., 2011) zugrunde.

Folgende Faktoren zeigten einen signifikanten Einfluss auf den *weighted Recall*:

- Alle drei Parameter P_L, P_T und P_L und ihre gegenseitigen Wechselwirkungen. Eine dreiwertige Interaktion zu betrachten ist nicht notwendig.
- Die beiden Variablen *representation* und *case*, die die Darstellung des Textes beschreiben, und ihre Interaktion.
- Die Interaktion zwischen *representation* und P_L und zwischen *representation* und P_T.
- Die Länge des segmentierten Satzes in *Zeichen*. Der Effekt ist mit $(-3.3 \pm 1.3) \cdot 10^{-4}$ klein. Auf 100 Zeichen sinkt der *weighted Recall* um etwa 0.03 ab. Aber mit einem p-Wert von 0.023 kann der Effekt dennoch als recht sicher gelten. Abbildung 2.21 zeigt den Effekt im Überblick.

Vergleicht man die beiden Abbildungen 2.14 und 2.22, so erkennt man eine hohe Korrelation zwischen den Effekten, die die jeweiligen Parameter in beiden Datensätzen auf die jeweilige Messgröße ausüben. Der auffälligste Unterschied ist, dass $P_L =$ `longest` für den händisch erstellten Goldstandard noch weniger gute Ergebnisse zeigt als für die

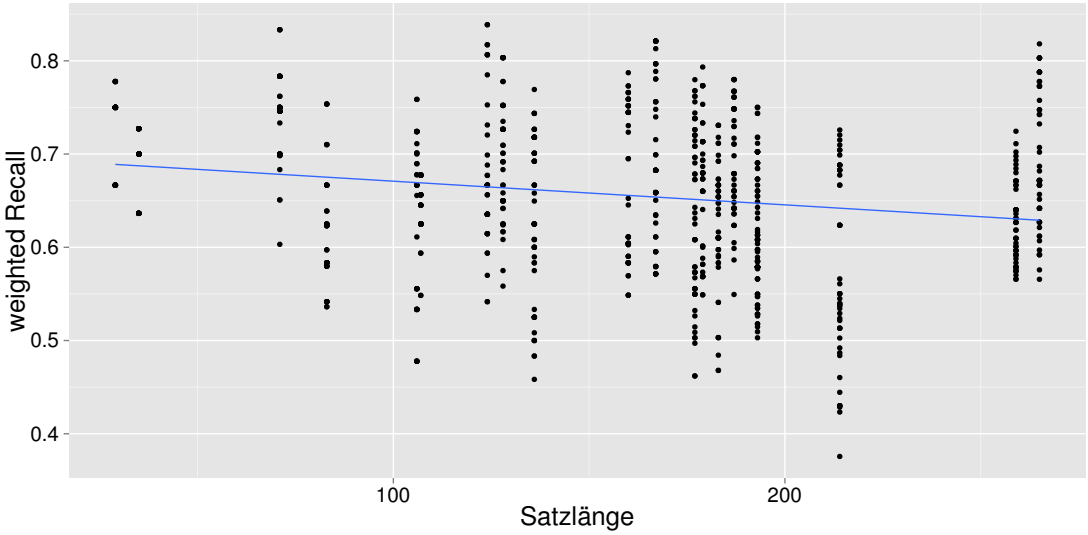


Abbildung 2.21: Der Satzlängeneffekt. Für *weighted Precision* und *weighted f* lassen sich analoge Bilder zeichnen. Die Gerade entspricht einer einfachen Regression, hat also nur Übersichtscharakter. Die senkrechten Streifen sind auf die konstanten Satzlängen der 20 Testsätze zurückzuführen.

Erkennung der Leerzeichen als Segmentgrenze. Dies ist nicht überraschend. Morpheme sind tendenziell kürzer als (orthographische) Wörter. Daher ist eine Strategie, die längere Segmente bevorzugt, hier naturgemäß nicht so erfolgreich.

Diese Übereinstimmung zwischen den auf einem vollständigen Goldstandard beruhenden und einer nur auf den Leerzeichen aufbauenden Analyse ist im Nachhinein auch ein Argument für die Gültigkeit der für die anderen dort untersuchten Sprachen Englisch und Türkisch abgeleiteten Ergebnisse.

Nach dieser Konsistenzprüfung der Methode und der Daten betreten wir mit der Analyse der Evaluationsmaßes *weighted Precision* und *weighted f* Neuland.

Die signifikanten Parameter unterscheiden sich in beiden Fällen leicht von den oben dargestellten Ergebnissen für *weighted Recall*.

- Im Falle der *weighted Precision* verlieren die Satzlänge und die Interaktionen zwischen P_F und P_T und zwischen *representation* und P_T ihre Signifikanz.
- Im Falle des *weighted f measure* verliert nur die Wechselwirkung zwischen *representation* und P_T ihre Signifikanz.

Interessant scheint hier vor allem die scheinbare Asymmetrie des Längeneffektes zwischen *weighted Recall* und *weighted Precision*. Ob dies ein stabiles Feature ist, bleibt eine spannende Frage für weitere Forschungen. Wenn sich stabile Resultate zeigen, dass der *weighted recall* im Gegensatz zur *weighted precision* zum Satzende abnimmt, wäre das ein erstaunliches Ergebnis, das dringend einer linguistischen Erklärung bedürfte. Ein

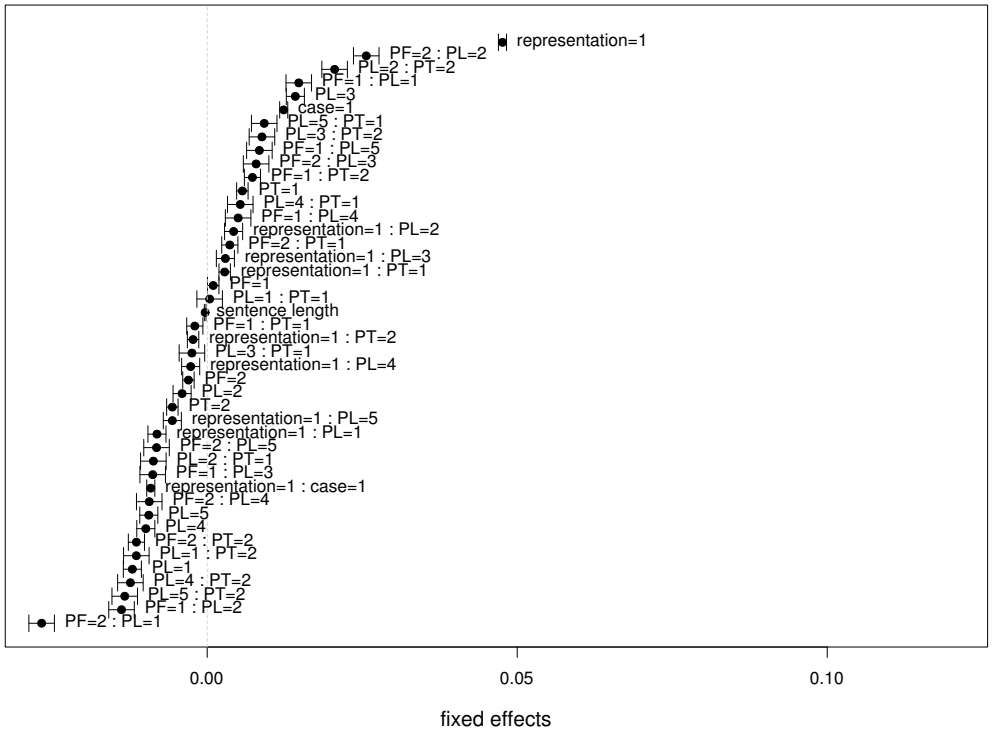


Abbildung 2.22: Graphische Darstellung des Einflusses der Parameter auf den *weighted Recall*. Die Fehlerbalken geben den Standardfehler an. Der Achsenabschnitt (Intercept) ist so weit im positiven Bereich, dass er nicht in die Graphik übernommen wurde.

einziger Befund mit einem p-Wert von über 0.02 ist sicher noch kein ausreichender Beleg für solch ein unerwartetes Phänomen.

Der Parametersatz mit maximalem *weighted* f ist nicht vollkommen identisch mit dem im vorigen Abschnitt ermittelten, in dem es um die Erkennung der Leerzeichen als Morphemgrenzen ging. Dort hatte sich die optimale Kombination $P_L =$ `combined`, $P_F =$ `sum` und $P_T =$ `tree_sum` ergeben. Hier nun schneidet $P_T =$ `tree_none` mit[77] $f_w = 0.787 \pm 0.003$ am besten ab. Der bisherige Optimalwert $P_T =$ `tree_sum` allerdings bleibt mit $f_w = 0.786 \pm 0.003$ nur marginal darunter. Dieser nicht signifkante Unterschied wäre im Anwendungsfall sicherlich zu vernachlässigen. Gerade aus der Anwendungsperspektive ist es ein beruhigender Hinweis auf die Stabilität des optimalen Parametersatzes, dass sich für die beiden unterschiedlichen Datensätze und Evaluationsmaße so übereinstimmende Ergebnisse zeigen.

Abbildung 2.23 zeigt nun das Zusammenspiel von *weighted Recall*, *Precision* und f in

[77] Der angegebene Fehler ist der Standardfehler, gemessen über die 20 Testsätze.

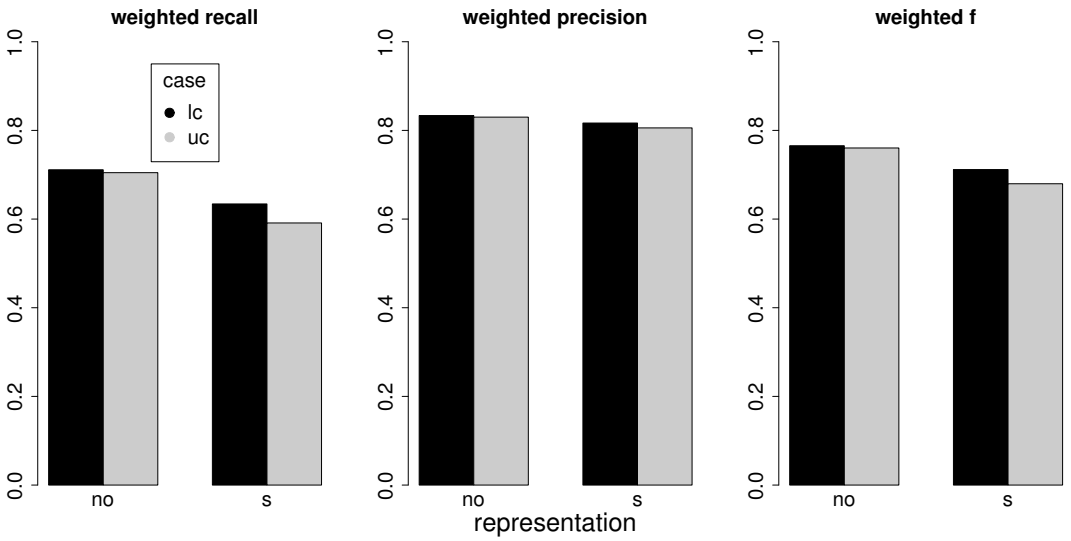


Abbildung 2.23: *weighted Recall, Precision* und f in den 4 Darstellungen des Textes. Vergleiche auch Abbildung 2.11.

den 4 aus den Kombinationen von *case* und *representation* gebildeten Darstellungen des Textes. Einige Charakteristika fallen auf:

- Es gibt wieder eine grundlegende Übereinstimmung mit den in Abbildung 2.11 dargestellten Ergebnissen der *Performanz* auf dem Datensatz aller Leerzeichen.
- *weighted Precision* übersteigt *weighted Recall*.
- Generell ist die Performanz mit Leerzeichen (*representation*=`no`) besser als ohne (*representation*=`s`).
- Dieser Effekt ist wesentlich stärker in Bezug auf den *weighted Recall* als in Bezug auf die *weighted Precision*. Das heißt, die Morphemgrenzen verlieren mit dem Leerzeichen an Sichtbarkeit, aber dies beeinträchtigt kaum die Treffsicherheit der vorgeschlagenen Grenzen.
- Die Variable *case* macht einen kleineren Unterschied. `lc` schneidet besser ab als `uc`. Auch hier ist der Einfluss auf den *weighted Recall* größer. Dies vor allem in der Darstellung ohne Leerzeichen (*representation*=`s`). Das heißt, ohne Leerzeichen wirkt sich eine Verbesserung der Statistik durch die Nivellierung der Groß- und Kleinschreibung besonders positiv aus.

Erste Erkenntnisse zum letzten Punkt wurden bereits auf den Seiten 82 und 88 dargestellt. Es kann damit als gesichert gelten, dass im vorliegenden deutschen Korpus die originale Schreibweise (*case*=`uc`) schlechter abschneidet. Auf den ersten Blick ist das kontraintuitiv: Für einen menschlichen Leser macht Groß- und Kleinschreibung einen

deutschen Text gerade nach Löschung der Leerzeichen lesbarer. Dies steht dem Verhalten des Algorithmus genau entgegen. Dem Computer aber gelten das große `A` und das kleine `a` aber als zwei völlig verschiedene Zeichen, während für einen Sprecher beide eindeutig zu einer gemeinsamen Kategorie gehören. In diesem Sinne wird der Algorithmus durch die Normalisierung auf Kleinschreibung mit sprachlichem Wissen versorgt.

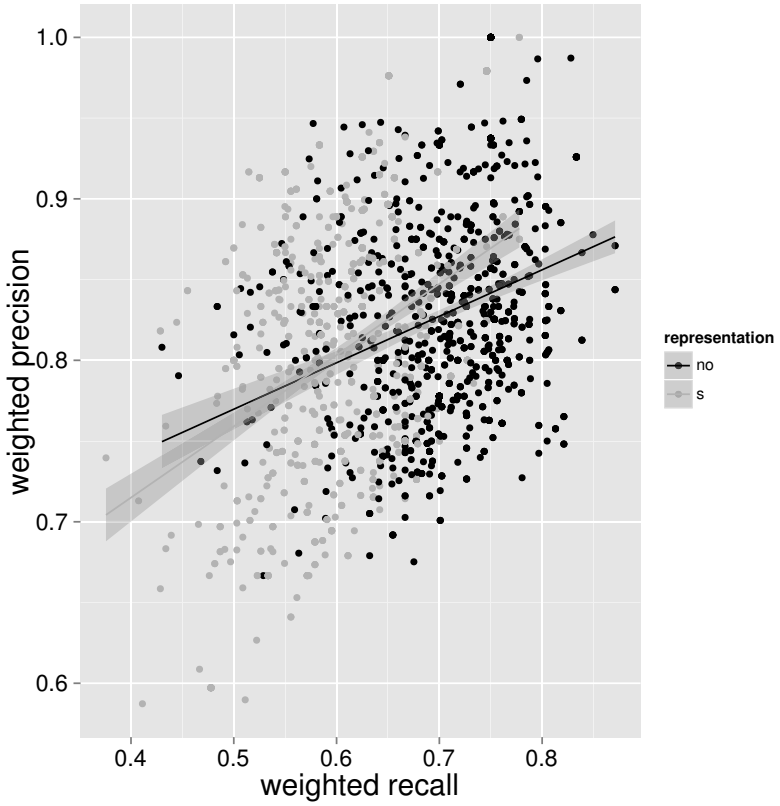


Abbildung 2.24: *weighted Recall* und *Precision* als Scatterplot.

Es ist zu erwarten, dass *weighted Precision* und *Recall* gegenläufige Tendenzen zeigen. Veränderungen, die eine Erhöhung des *Recalls* zur Folge haben, resultieren meist in einer Abnahme der *Precision*. Der naive Algorithmus, der zwischen allen *Zeichen* Segmentgrenzen setzt, hat maximalen *Recall*, da so alle echten Grenzen gefunden werden. Die Precision liegt allerdings weit darunter, da längst nicht nach jedem *Zeichen* eine Segmentgrenze liegt. Genau diese Gegenläufigkeit ist die Motivation, die hinter der Definition von *weighted f* steht. *f* ist nur dann in der Nähe von 1, wenn dies für die beiden Grundmaße *(weighted) Recall* und *Precision* gleichermaßen gilt.

Eine Anmerkung: Dieses Argument bezieht sich auf Änderungen im Algorithmus, nicht auf die Struktur der Daten. Das heißt: Eine Strategie, die einen exzellenten *Recall* liefert neigt stark zu einer geringen *Precision*. Ein Satz aber, der gut segmentierbar ist, hat meist für beide Maße hohe Werte. Abbildung 2.24 zeigt diesen Effekt.

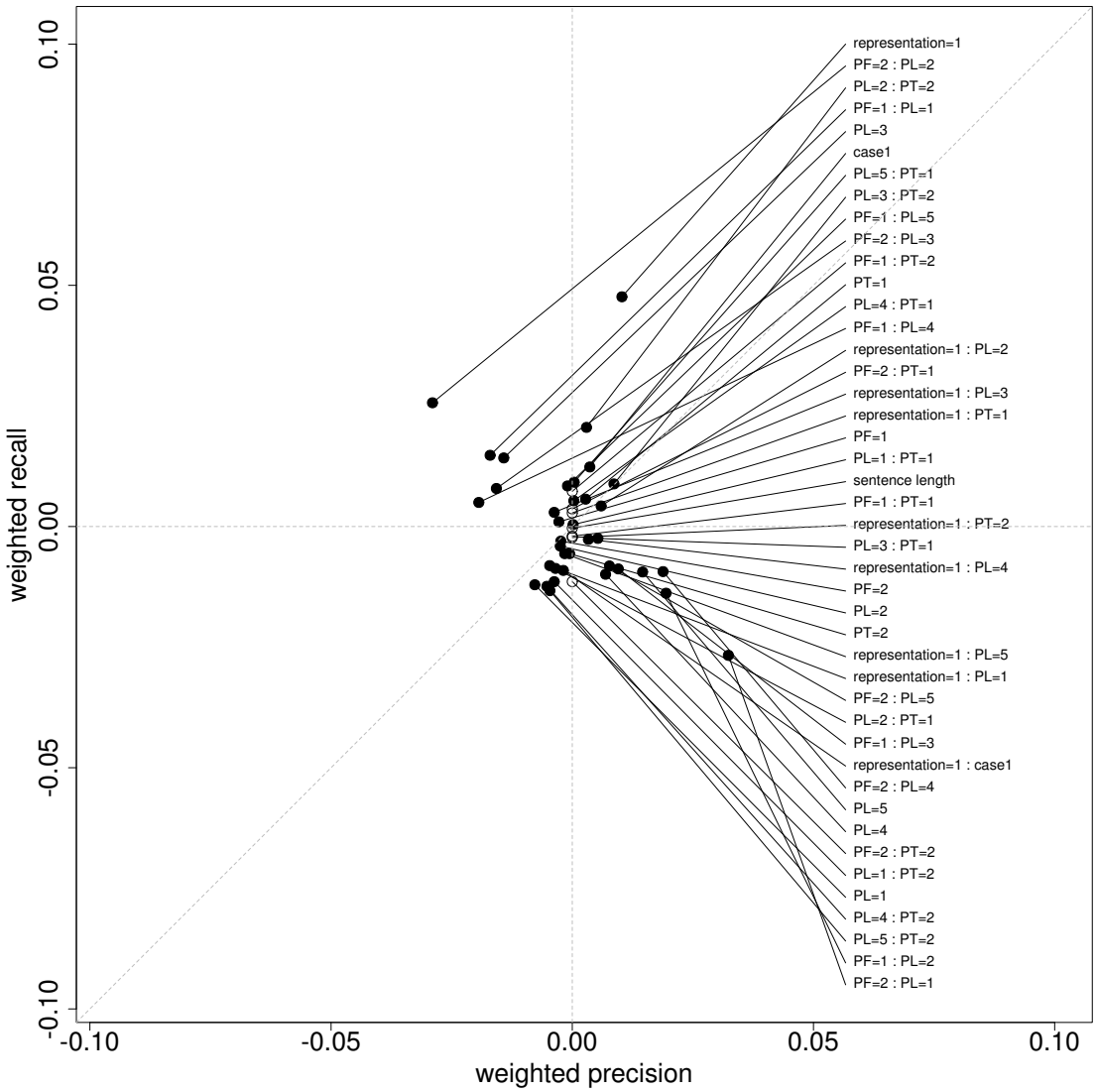


Abbildung 2.25: Vergleich der Effekte der Parameter auf *weighted Recall* und *weighted Precision*. Die mit offenen Kreisen gekennzeichneten Parameterwerte haben keinen signifikanten Einfluss auf die *weighted Precision*.

Abbildung 2.25 zeigt den Einfluss der verschiedenen Parameterstellungen jeweils auf *weighted Recall* und *Precision*. In vielen Fällen erkennt man die erwartete Gegenläufigkeit. So ist $P_L =$ `longest` günstig für die *Precision* und ungünstig für den *Recall*. Dies ist einsehbar: Wenn lange *Segmente* bevorzugt werden, sind diese zwar oft korrekt, viele *sprachliche Segmente* werden aber verpasst.

Ähnlich nachvollziehbar ist der sehr starke Effekt der Kombination aus $P_F =$ `sum` und $P_L =$ `longest`, der genau die gegenläufigen Tendenzen zeigt. Die Summierung der Scores bis zum Satzende hebt die Bevorzugung der längsten Segmente gerade eben auf, da sich

immer dieselbe Gesamtlänge ergeben muss, wenn man von Leerzeichen absieht, die die Bilanz ein wenig verändern können. Dieses Aufheben des Effektes von P_L = `longest` durch P_F = `sum` zeigt sich in der gegenüberliegenden Lage der beiden Punkte (P_L = `longest` und P_L = `longest` + P_F = `sum`) rechts unten und links oben.

Aus Abbildung 2.23 ist uns die Wirkung von *representation*=`s` bereits bekannt. Sowohl *weighted Recall*, als auch *Precision* sind hier reduziert, wobei aber der *Recall* wesentlich stärker reduziert ist.

Es zeigen sich aber auch Effekte in Abbildung 2.25, die nicht so leicht erklärbar sind. Wieso wirkt sich die Kombination aus *representation*=`s` und P_L = `longest` negativ aus, sowohl auf *Recall*, als auch auf *Precision*? Wieso ist die Kombination P_F = `sum` und P_L = `forward` eher günstig für den *Recall*, aber ungünstig für die *Precision*? Gerade diese auffälligen, aber nicht unmittelbar erklärlichen Effekte könnten sich in der Zukunft als der Ansatzpunkt erweisen, weitergehende Erkenntnisse aus den Daten zu ziehen.

Ein interessanter wie erfreulicher Punkt ist, dass der Haupteffekt von P_L = `combined` sowohl für die *weighted Precision*, als auch für den *weighted Recall* günstig ist. Dies deutet wie alle bisherigen empirischen Ergebnisse darauf hin, dass die konsequente Ausnutzung der vollen verfügbaren Frequenzinformation in Form der beiden *predictability changes* ein Optimum darstellt. Dies ist eine weitere Bestätigung *a posteriori* für den Ausgangspunkt des gesamten Algorithmus. Auch aus Anwendungsperspektive ist es günstig, dass ein und dieselbe Verfahrensvariante konsequent optimale Ergebnisse liefert.

Wenn dies sich für mehr Korpora und mehr Sprachen als ein stabiles Feature erweisen sollte, könnte man den überwachten Aspekt des Algorithmus entfernen und ein tatsächlich maximal unüberwachtes Verfahren erhalten. Wie oben erwähnt (2.3) kann man sich auf den Standpunkt stellen, dass die Parameter $P_{LFT(4)}$ frei variabel sind und über die hier besprochene Evaluation der Ergebnisse optimiert werden. Kämen derartige Evaluationen konsistent zum selben Ergebnis, kann der optimale Parametersatz festgeschrieben werden.

Nun kennen wir die Wirkung der verschiedenen Parameterstellung auf *weighted Recall* und *Precision* und ihr Zusammenspiel. *weighted f* fasst beide Evaluationsmaße in ein einziges zusammen. Abbildung 2.26 zeigt den Einfluss der Parameter auf *weighted f*.

Sogleich springt wieder die hervorragende Performanz von P_L = `combined` ins Auge. Dieser Parameter hat von allen die größte Wirkung. Auch der Vorteil von *case* = `lc` gegenüber `uc` und die starke Wechselwirkung dieses Parameters mit *representation* bestätigt sich.

Die Asymmetrie zwischen P_L = `forward` und P_L = `backward` begegnet uns hier ebenfalls wieder. Wieder ist es `forward`, das besser abschneidet. Und wie schon in den Abbildungen 2.14 bis 2.16 liegt auch hier die Methode P_L = `children`, die die Zahl der Kinder unterhalb eines Knotens maximiert, zwischen diesen beiden.

Wenden wir uns nun einem Parameter zu, der bisher vernachlässigt wurde. Auf Seite 67 wurde er unter dem Namen P_4 bereits kurz erwähnt. Dieser Parameter entscheidet darüber, welche Kombination an Kindern gewählt wird, wenn für einen Vaterknoten verschiedene Möglichkeiten existieren. Der Einfluss dieses Parameters liegt um etwa eine Größenordnung unter der Wirkung von zum Beispiel P_L. Auch der Einfluss von P_F und P_T ist 4 bis 5 mal größer. Daher ist er aus Anwendungssicht nicht besonders interessant.

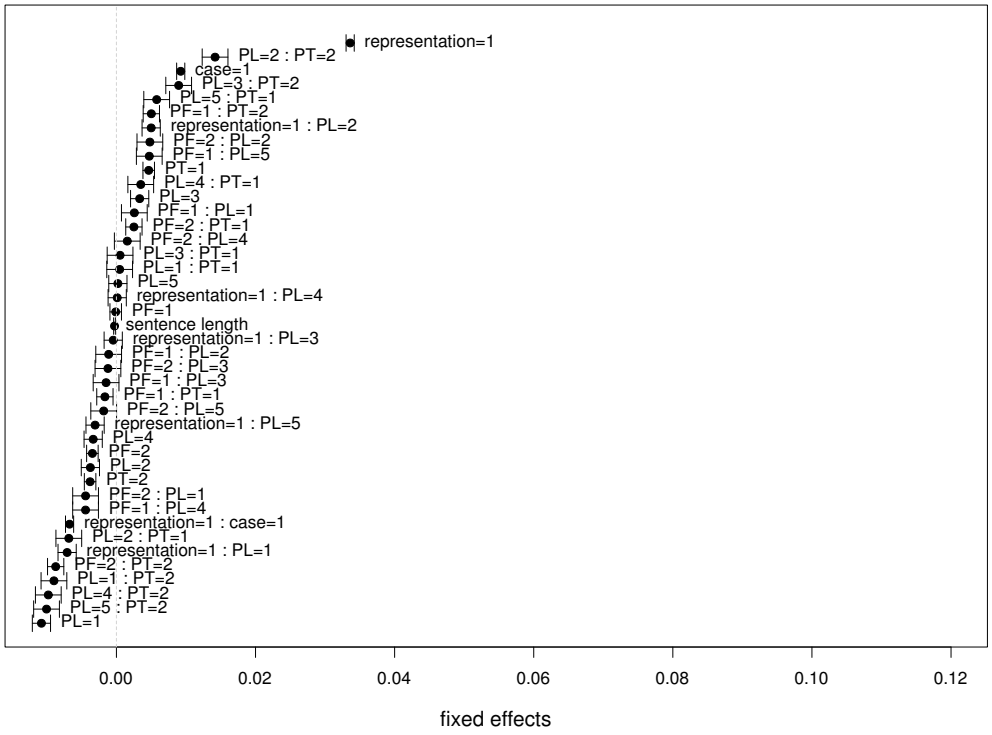


Abbildung 2.26: Graphische Darstellung des Einflusses der Parameter auf *weighted* f. Die Fehlerbalken geben den Standardfehler an. Der Achsenabschnitt (Intercept) ist so weit im positiven Bereich, dass er nicht in die Graphik übernommen wurde. Zum besseren Vergleich stimmt die x-Achse mit dem in Abbildung 2.22 für den *weighted Recall* verwendeten Bereich überein.

Ihn dennoch auszuwerten gibt aber noch einen weiteren Blick auf das Zusammenspiel von Daten und Algorithmus frei.

Es folgt nun eine Auflistung der 13 möglichen Verfahren, die hier getestet wurden. Sei f der Vaterknoten des Teilbaumes, den es zu bewerten gilt. lk ist sein linker Kindknoten, rk sein rechter. Wie bisher auch bezeichnen $N_{(T)}(f)$, $N_{(T)}(lk)$ und $N_{(T)}(rk)$ die Häufigkeiten dieser Zeichenketten im Trainingskorpus.

Die Motivation für die einzelnen Varianten war recht unterschiedlich: Einige sind nur ein Konsistenztest, das heißt, sie wurden nur untersucht um zu prüfen, ob sie so schlecht abschneiden wie erwartet. Andere beinhalten nur die Frequenz der einzelnen Zeichenketten. Dies ist vor allem interessant im Vergleich zur dritten Gruppe, die auf dem hier zentralen Begriff des *predictability change* beruhen. So kann man Einsicht in die Frage gewinnen, ob auch an dieser Stelle die Verwendung der sich verändernden Vorhersag-

barkeiten den reinen Frequenzen überlegen ist. Ein solches Ergebnis würde durchaus die Überlegung motivieren, ob dies nur eine Begleiterscheinung des Algorithmus selbst ist oder eine Eigenschaft des erzeugenden Systems der Sprache bzw. ob menschliche Segmentierungsstrategien ähnliche Eigenschaften haben.

Im einzelnen wurden folgende Strategien untersucht:

$P_4 =$ `forward_pred` Welcher Anteil der Vorkommen des linken Kindes setzt sich zum Gesamtstring, also zu f fort?

$$I_{P4} = -\frac{N(f)}{N(lk)}$$

Bevorzugt werden Konstellationen, in denen sich der Gesamtstring gut aus dem linken Kind vorhersagen lässt.

$P_4 =$ `backward_pred` Welcher Anteil der Vorkommen des **rechten** Kindes setzt sich nach links zu f fort, folgt also auf ein Vorkommen von lk?

$$I_{P4} = -\frac{N(f)}{N(rk)}$$

Bevorzugt werden Konstellationen, in denen sich der Gesamtstring gut aus dem rechten Kind (rückwärts) vorhersagen lässt.

$P_4 =$ `pred` Die Kombination aus den beiden vorhergehenden:

$$I_{P4} = -\left(\frac{N(f)}{N(lk)} + \frac{N(f)}{N(rk)}\right)$$

$P_4 =$ `frequent_frequent` Bevorzugt werden Konstellationen, in denen eines der Kinder besonders häufig ist:

$$I_{P4} = -\max(N(lk), N(rk))$$

$P_4 =$ `rare_frequent` Bevorzugt werden Konstellationen, in denen **das seltenere** Kind besonders häufig ist:

$$I_{P4} = -\min(N(lk), N(rk))$$

$P_4 =$ `frequent_rare` Bevorzugt werden Konstellationen, in denen **das häufigere** Kind besonders selten ist:

$$I_{P4} = \max(N(lk), N(rk))$$

$P_4 =$ `rare_rare` Bevorzugt werden Konstellationen, in denen **das seltenere** Kind besonders selten ist:

$$I_{P4} = \min(N(lk), N(rk))$$

Diese Möglichkeit ist eher eine Konsistenzprüfung. Es kann erwartet werden, dass sie zu negativen Ergebnissen führt.

P_4 = `middle_forward` Bevorzugt werden Konstellationen, in denen das linke Kind einen besonders großen *forward predictability change* hat.

$$I_{P4} = D^{+}(lk)$$

Dies ist der *forward predictability change* zwischen den Kindern. Je kleiner er ist, desto stärker fällt die Vorhersagbarkeit ab. Daher das positive Vorzeichen.

P_4 = `middle_backward` Bevorzugt werden Konstellationen, in denen das **rechte** Kind einen besonders großen ***backward*** *predictability change* hat.

$$I_{P4} = D^{-}(rk)$$

Wieder geht es um die Stelle zwischen den beiden Kindern.

P_4 = `all_middle_drops` Bevorzugt werden Konstellationen, in denen die logarithmische Summe beider *predictability changes* zwischen den Kindern minimal (also extrem) ist.

$$I_{P4} = \log\left(D^{+}(lk)\right) + \log\left(D^{-}(rk)\right)$$

P_4 = `all_drops` Erweiterung: Bevorzugt werden Konstellationen, in denen die logarithmische Summe aller 4 *predictability changes* der Kinder minimal (also möglichst extrem) ist.

$$I_{P4} = \log\left(D^{+}(lk)\right) + \log\left(D^{-}(lk)\right) + \log\left(D^{+}(rk)\right) + \log\left(D^{-}(rk)\right)$$

P_4 = `forward_drops` Bevorzugt werden Konstellationen, in denen die logarithmische Summe beider *forward predictability changes* minimal ist.

$$I_{P4} = \log\left(D^{+}(lk)\right) + \log\left(D^{+}(rk)\right)$$

P_4 = `backward_drops` Bevorzugt werden Konstellationen, in denen die logarithmische Summe beider ***backward*** *predictability changes* minimal ist.

$$I_{P4} = \log\left(D^{-}(lk)\right) + \log\left(D^{-}(rk)\right)$$

Abbildung 2.27 zeigt die Wirkung der verschiedenen Werte von P_4 auf *weighted* f. Man kann drei Gruppen erkennen, die sehr ungünstigen Möglichkeiten, die mittleren und die guten. Ungünstig ist es, die Seltenheit von *Segmenten* zu belohnen (P_4 = `rare_rare|frequent_rare`). Dies war erwartbar und stellt einen Konsistenztest dar.

Ähnlich schlecht schneidet P_4 = `pred` ab, die Kombination aus `forward_pred` und `backward_pred`. Die Vorhersagbarkeit des Gesamtstrings aus dem rechten und dem linken Kind gleichzeitig zu maximieren, ist offenbar ein schlechter Kompromiss, vermutlich, da so insgesamt wieder seltenere *Segmente* bevorzugt werden: Unter der Voraussetzung, dass zwischen den beiden Kindsegmenten in beiden Richtung ein starker

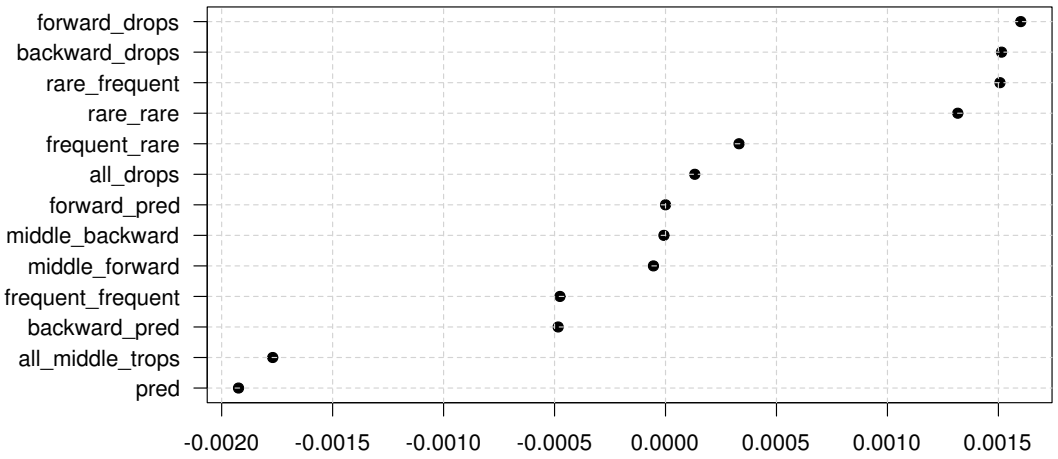


Abbildung 2.27: Der Einfluss von P_4. Die Y-Achse listet seine 13 möglichen Einstellungen auf. Die x-Achse zeigt ihren jeweiligen Einfluss auf *weighted f* relativ zum (willkürlich ausgewählten) Referenzniveau `backward_pred`.

predictability drop vorliegt ist es schwer vorstellbar, dass dennoch aus beiden Teilen der Gesamtstring gut vorhersagbar ist.

Die Varianten $P_4 =$ `frequent_frequent|rare_frequent` bevorzugen häufige Kinder. Dass diese Verfahren im Mittelfeld liegen ist nicht unerwartet. Einer der Grundgedanke der *Morphologischen Induktion* insgesamt ist schließlich, dass man *Segmente* an ihrer überproportionalen Häufigkeit erkennen kann.

Es folgen linguistisch interessantere Varianten. Bemerkenswert ist der große Unterschied zwischen $P_4 =$ `middle_forward` und `middle_backward`. Dieser findet sich wieder in `forward_drops` und `backward_drops`. Es ist sogar ein wenig besser ausschließlich den *backward predictability change* zwischen den beiden Kindern zu verwenden, als alle vier *predictability changes* beider Kinder. Zwei Dinge sind hierzu anzumerken.

Einerseits ergeben sich vertauschte Rollen von *forward* und *backward predictability change* im Vergleich zu P_L. Dieses sehr klare Ergebnis scheint ein lohnendes Ziel weiterer Forschung. Andererseits ist es hier ausgeschlossen, dass die *forward-backward*-Asymmetrie auf die Verarbeitungsrichtung zurückzuführen ist. Aufgrund des rekursiven Algorithmus, der sich von links nach rechts durch den Satz arbeitet, scheint dies im Falle von P_L zumindest vorstellbar. Hier geht es aber jeweils nur um die gleichzeitige Beurteilung zweier aufeinanderfolgenden *Segmente* an einer bestimmten Stelle im Satz.

Insgesamt ist es gelungen, reiche Strukturen aufgrund eines sehr kleinen Goldstandards aufzudecken. Die Eigenschaften, die sich ohne weiteres vorhersagen lassen, finden sich in den Daten. Dies verleiht den Ergebnissen insgesamt Plausibilität. Die Fragen, die sich ergeben, scheinen fruchtbare Ansatzpunkte für weitere Untersuchungen. Weitere Variationen für P_L, P_T, P_F und P_4 sind vorstellbar.

2.6.4 Manuelle Evaluation eines Querschnitts der entstehenden Segmente

Nun ist einiges bekannt über die Performanz des Algorithmus in Bezug auf die gesetzten Segmentgrenzen. Eine wichtige Lücke gilt es aber noch zu schließen:

In einem dritten Evaluationsschritt soll über die Betrachtung der Segmentgrenzen hinausgegangen werden. Die *Segmente* sind nun Thema. Es wird untersucht

- welchen linguistischen Status die korrekt erkannten *sprachlichen Segmente* haben.
- in welchem Ausmaß Fehler auftreten.
- welche Arten von Fehlern in welcher Verteilung auftreten.
- in welcher Umgebung bzw unter welchen Umständen Fehler auftreten.
- aus welchem Grund Fehler auftreten.

In einem letzten Gedankenschritt wird eine mögliche Lösungsstrategie skizziert.

Die folgenden Betrachtungen werden insgesamt einen sehr qualitativen und ausschnittartigen Charakter haben. Leider gibt es keinen Goldstandard, der eine vollständigere Evaluation erlauben würde. Der in Abschnitt 2.6.3 verwendete ist hierfür viel zu klein. Auch enthält er nur eine Segmentierung in *minimale sprachliche Segmente*, größere Einheiten fehlen. Darüber hinaus ist er nur für eine der drei Sprachen verfügbar. Es kann hier also nur darum gehen, wenigstens einen Überblick über die Struktur der Ergebnisse zu bekommen.

Datengrundlage der Untersuchung war die durch *representation* = `s` und *case* = `lc` gekennzeichnete Textversion.[78] Aus Performanzgesichtspunkten ist dies eine suboptimale Wahl, da die Textversionen mit Leerzeichen erheblich besser abschneiden. Hier soll es aber gerade nicht um die genaue Performanz unter verschiedenen Parameterwerten oder um deren Wirkung gehen. All dies war in den bisherigen empirischen Untersuchungen bereits Thema. Hier geht es um eine qualitative Einordnung der Ergebnisse, insbesondere der auftretenden Fehler. In diesem Zusammenhang kann man argumentieren, dass es günstiger ist, eine Textvariante mit eher mehr als weniger Fehlern zu untersuchen.

Verwendet wurden jeweils die Ergebnisse für den Parametersatz P_L = `combined`, P_T = `tree_sum`, P_F = `none` (und P_4 = `middle_forward`). Dies ist wiederum nicht der Parametersatz mit dem höchsten *weighted f* für diese Textversion, sondern liegt etwa einen Prozentpunkt darunter. Dies ist aus dem oben erwähnten Grund nicht von Belang und es ist auch nicht zu erwarten, dass die auftretenden Fehler von Parametersatz zu Parametersatz allzu stark schwanken.

Für alle drei untersuchten Sprachen wurde eine sortierte Frequenzliste der entstehenden Segmente erstellt. Aus dieser Liste wurden drei mal zehn Segmente ausgewählt:

- die 10 häufigsten Segmente
- die ersten 10 mit Häufigkeit 5

[78] Präzise gesagt wurden für diese Untersuchung nicht nur die Leerzeichen, sondern auch die Satzzeichen entfernt.

- eine Auswahl aus den Segmenten, die nur ein einziges Mal vorkamen (*hapax legomena*). Diese Segmente sind erwartbar am zahlreichsten. Da bei alphabetischer Ordnung leicht unrepräsentative Reihen entstehen, wurde für Deutsch nur jedes 5. Segment in der Liste berücksichtigt, und für Englisch und Türkisch nur jedes 20.

Nun gilt es diese 3×30 Segmente zu bewerten. Eine naheliegende Einteilung wäre schlicht in *richtig* und *falsch*, je nachdem ob das vom System vorgeschlagene *Segment* mit einem *sprachlichen Segment* zusammenfällt, oder nicht. Dies ist durchaus ein etabliertes Vorgehen, das klassische Goldstandardverfahren beruht darauf. Nun greift aber die Zweiteilung in richtig und falsch oftmals zu kurz. Dies ist schon an oben verwendeten kleinen Goldstandard erkennbar, in dem etwa 15% der Segmente nicht von allen drei Experten gesetzt wurden. Entsprechend unterschiedlich sollten verschiedene Fehler gewichtet werden. Um diesen Weg konsequent zu gehen, wäre ein ausreichend großer, manuell von verschiedenen Experten erstellter, Goldstandard für alle drei Sprachen erforderlich.

Ein anderer Aspekt der Fehler ist vor allem in Bezug auf mögliche Anwendungen vielleicht auch interessanter. Daher habe ich mich entschieden, die Segmente folgendermaßen zu klassifizieren. Für jedes Vorkommen eines *Segmentes* s habe ich eine von vier möglichen Bewertungen vergeben. Dabei zog ich nur mein eigenes linguistisches Wissen in Betracht. Die Theorieabhängigkeit und graduelle Sicherheit *sprachlicher Segmente* wird dabei nicht berücksichtigt. Für eine explorative Untersuchung ist dies ausreichend.

1. Das Segment s ist korrekt, d.h., es fällt mit einem *sprachlichen Segment* zusammen. In diesem Fall wurde gezählt, wie viele *minimale sprachliche Segmente* das betrachtete Segment enthält.

2. Für alle übrigen Fälle wurden drei Möglichkeiten unterschieden:

   a) s entsteht durch Übersegmentierung. Das Segment $s =$`the` in `the|ater` ist zwar eindeutig falsch, es umschließt aber keine Grenze eines *sprachlichen Segments*. Es können auch beide Segmentgrenzen eine Übersegementierung darstellen.

   b) s ist untersegmentiert. Seine beiden Grenzen fallen mit Grenzen *sprachlicher Segmente* zusammen. $s =$ `youabout` ist zwar kein *sprachliches Segment*, dennoch besteht diese Zeichenkette aus zwei vollständigen englischen Wörtern.

   c) Alle anderen Fälle: s umschließt mindestens eine Grenze eines *sprachlichen Segmentes* und hat selber mindestens eine Grenze, die nicht mit einer Grenze eines *sprachlichen Segmentes* übereinstimmt: $s =$`ua` in `yo|ua|bout` wäre ein Beispiel.

Fehler der Kategorie 2a können als relativ leicht gelten. Einerseits verhindern sie nicht die Erkennung *sprachlicher Segmente* auf einer höheren Ebene, andererseits könnten sie theoretisch durch eine einfache Beseitigung des fraglichen Segments behoben werden. Sie zerstören also nicht die tatsächliche Struktur des Satzes, sondern erweitern sie nur durch *falsche Segmente*.

Fehler der Kategorie 2b gibt es im Grunde nur für die längeren Segmente der Frequenz 1. Die vergleichsweise geringe Verlässlichkeit dieser Segmente kann relativ zuverlässig an den kleinen Frequenzzahlen abgelesen werden, auf denen ihr Segmentstatus beruht. Das Einführen einer Mindestfrequenz würde sie bereits weitgehend beseitigen.

Am unangenehmsten sind Fehler der Kategorie 2c. Sie überschneiden sich mit tatsächlichen *sprachlichen Segmenten*. Derartige Fehler sind weder durch Vereinigung oder Teilung von Segmenten alleine zu beseitigen, sie liegen quer zur tatsächlichen morphologischen Struktur. Die Tatsache, dass diese Art Fehler im untersuchten Datenquerschnitt kaum eine Rolle spielt, kann als sehr ermutigend beurteilt werden.

Die Ergebnisse der Analyse sind in den Tabellen 2.2, 2.3 und 2.4 zusammengefasst. Der Anteil von 80 bis 85% korrekten Segmente kann wegen des gewählten Datenquerschnitts nicht direkt als *Precision* interpretiert werden, setzt aber einen gewissen Rahmen. Da die Gesamtmenge aller in einem Text enthaltenen *sprachlichen Segmente* schwer zu ermitteln ist, kann kein dem *Recall* auch nur verwandter Begriff angegeben werden. Von dem hier verwendeten Algorithmus kann aber auch kaum eine vollständige Zerlegung eines Satzes erwartet werden. Dazu müsste jeder Satz mindestens einmal auch im Trainingskorpus vorkommen. Daher wäre die Angabe eines *Recall* auch nur von sehr begrenztem Wert.

Für einzelne Segmente ist die *Precision* aber durchaus berechen- oder zumindest abschätzbar. So kommt die am häufigsten als *Segment* vorgeschlagene Zeichenkette `en` im deutschen Testkorpus 917 mal vor. 195 mal wurde es als Segment klassifiziert, 194 Fälle davon wurden ausgewertet.[79] 177 erwiesen sich als korrekt, 15 als inkorrekt. Dies entspricht einer *Precision* von 0.91. Den *Recall* kann man aus dem verwendeten Goldstandard abschätzen. Dort ist ein Anteil von 0.68 ± 0.10 der Vorkommen der Zeichenkette `en` ein sprachliches Segment.[80] Dies lässt für das Gesamtkorpus zwischen 530 und 707 *sprachliche Segmente* `en` erwarten. Daraus folgt ein *Recall*[81] von 0.29 ± 0.05. Man kann versuchsweise annehmen, dass sich für viele *Segmente* ähnliche Verhältnisse von *Recall* und *Precision* ergeben. Die hohe *Precision* ist sicherlich positiv. Der geringe *Recall* kann zwar als Manko gesehen werden. Er kann aber auch als Reflexion der Beobachtung gesehen werden, dass viele Oberflächenformen so häufig in einem Stück vorkommen, dass ihre Aufteilbarkeit in mehrere *Segmente* mehr und mehr in den Hintergrund tritt. Ein deutsches Beispiel wäre das Segment `nebenbeibemerkt`. Es wird normgerecht in zwei Wörtern geschrieben, taucht aber so häufig zusammen auf, dass es die Funktion eines einzigen Adverbs übernimmt. Ein sehr ähnlicher Fall, der bereits zusammengeschrieben wird, wäre „kurzerhand“.

Welche Segmente werden korrekt erkannt? Es zeigt sich, dass erwartungsgemäß die häufigsten erkannten *sprachlichen Segmenten* entweder Vertreter grammatischer Morpheme sind, oder Funktionswörter. Im mittleren Frequenzbereich folgen im wesentlichen einmorphemige Inhaltswörter. Die *Hapax Legomena* sind meist Verknüpfungen mehrerer Morpheme.

[79] Bei den sehr häufigen Segmenten waren mit der Standardanwendung unter Linux (`grep`) nicht ohne weiteres alle Vorkommen zu finden. In diesen Fällen habe ich darauf verzichtet, diese Vorkommen auszuwerten.

[80] Die Schwankungsbreite bezeichnet das Konfidenzintervall eines Binomialtests.

[81] Es ergäbe sich ein f von 0.44 ± 0.05.

Rang	Häuf.	String	korrekt	falsch	min. Seg.	Fehler
1	195	en	177	17	1	12,0,5
2	189	e	140	48	1	48,0,0
3	169	er	142	25	1	23,0,2
4	117	und	111	5	1	5,0,0
5	95	die	94	1	1	1,0,0
6	92	der	84	6	1	6,0,0
7	74	n	10	62	1	62,0,0
8	70	ich	58	11	1	11,0,0
9	66	sch	0	62	0	62,0,0
10	62	war	59	3	1	3,0,0
183	5	zeit	5	0	1	-
184	5	wohn	5	0	1	-
185	5	werden	5	0	2	-
186	5	welch	5	0	1	-
187	5	stadt	5	0	1	-
188	5	spiel	5	0	1	-
189	5	son	0	5	-	5,0,0
190	5	soll	5	0	1	-
191	5	rat	2	3	1	3,0,0
192	5	öfter	5	0	2	-
877	1	zwölf	1	0	1	-
882	1	zweijahre	1	0	3	-
887	1	zuzeigen	1	0	3	-
892	1	zusein	1	0	2	-
897	1	zurverfügung	1	0	5	-
902	1	zunehmen	1	0	3	-
907	1	zumerst	0	1	-	0,0,1
912	1	zumachen	1	0	3	-
917	1	zujeder	0	1	-	0,1,0
		sum	924 78.7%	250 21,3%		241 (20.5%), 1 (0.09%), 8 (0.7%)

Tabelle 2.2: Deutsche Beispielsegmente und ihr linguistischer Status. Der Rang gibt den Platz des jeweiligen *Segmentes* in der sortierten Frequenzliste aller *Segmente* an. Die Spalte *Häuf*(igkeit) bezeichnet die Zahl der Vorkommen im Output. Die Spalten *korrekt* und *falsch* beinhalten meine Beurteilung der Vorkommen. Sie summieren sich nicht immer zur Gesamthäufigkeit auf, da aus technischen Gründen nicht immer alle *Segmente* beurteilt wurden. Die Spalte *min. Seg.* bezieht sich auf die Zahl der *minimalen Segmente* (≈Morpheme), die in der Zeichenkette maximal enthalten sind. *Fehler* schlüsselt die Fehler in der Form 2a,2b,2c auf.

Rang	Häuf.	String	korrekt	falsch	min. Seg.	Fehler
1	328	the	316	8	1	7 ,0,1
2	217	in	190	23	1	22,0,1
3	215	and	205	8	1	8,0,0
4	210	e	77	129	1	129,0,0
5	148	ing	121	26	1	25,0,1
6	147	a	103	36	1	36,0,0
7	135	of	133	2	1	1,0,1
8	110	it	87	21	1	17,0,4
9	108	on	46	61	1	61,0,0
10	107	ly	100	7	1	6,0,1
261	5	world	5	0	1	-
262	5	work	5	0	1	-
263	5	voice	5	0	1	-
264	5	visit	5	0	1	-
265	5	ve	3	2	1	2,0,0
266	5	ut	0	5	-	4,0,1
267	5	ur	0	5	-	5,0,0
268	5	understand	5	0	2	-
269	5	underst	0	5	-	0,0,5
270	5	tof	0	5	-	0,0,5
1411	1	ywent	0	1	-	0,0,1
1431	1	youthink	1	0	2	-
1451	1	youmean	1	0	2	-
1471	1	youabout	0	1	-	0,1,0
1491	1	yearsthis	0	1	-	0,1,0
1511	1	writer	1	0	2	-
1531	1	wor	0	1	-	1,0,0
1551	1	withhold	1	0	2	-
1571	1	willde	0	1	-	0,0,1
1591	1	whohe	0	1	-	0,1,0
		sum	1410 80,2%	349 19,8%		324 (18,4%), 3 (0,2%), 22 (1,3%)

Tabelle 2.3: Englische Beispielsegmente und ihr linguistischer Status. Der Rang gibt den Platz des jeweiligen *Segmentes* in der sortierten Frequenzliste aller *Segmente* an. Die Spalte *Häuf*(igkeit) bezeichnet die Zahl der Vorkommen im Output. Die Spalten *korrekt* und *falsch* beinhalten meine Beurteilung der Vorkommen. Sie summieren sich nicht immer zur Gesamthäufigkeit auf, da aus technischen Gründen nicht immer alle *Segmente* beurteilt wurden. Die Spalte („min. Seg.“) bezieht sich auf die Zahl der *minimalen Segmente* (≈Morpheme), die in der Zeichenkette maximal enthalten sind. *Fehler* schlüsselt die Fehler in der Form 2a,2b,2c auf.

Rang	Häuf.	String	korrekt	falsch	min. Seg.	Fehler
1	245	`e` (DAT)	201	31	1	31,0,0
2	227	`a` (DAT)	148	75	1	75,0,0
3	196	`i` (AKK)	49	138	1	138,0,0
4	152	`in` (GEN)	129	13	1	13,0,0
5	126	`bu` (dies)	123	0	1	-
6	118	`bir` (ein)	113	0	1	-
7	115	`lar` (PL)	114	1	1	1,0,0
8	115	`an` (PART)	69	43	1	43,0,0
9	107	`de` (in)	85	19	1	17,2,0
10	106	`ler` (PL)	103	0	1	-
560	5	`zorunlu` (gezwungen)	5	0	2	-
561	5	`yukarı` (oben)	5	0	1	-
562	5	`yol` (Weg)	5	0	1	-
563	5	`yönelik` (gerichtet auf)	5	0	3	-
564	5	`yönel` (zu)	5	0	1	-
565	5	`yükseldi` (wuchs)	5	0	2	-
566	5	`yazı` (Schreiben)	4	1	2	0,0,1
567	5	`yatırımcıların` (Investor|PL|GEN)	5	0	6	-
568	5	`yargı` (Urteil)	5	0	1	-
569	5	`yapacağı` (machen|FUT|AKK)	5	0	3	-
570	5	`yaklaş` (nähern)	5	0	2	-
2380	1	`ziyaretin` (Besuch|GEN)	1	0	2	-
2400	1	`üzerindekibaskıyı` (den darauf lastenden Druck)	1	0	7	-
2420	1	`zayıfla` (schwächen)	1	0	2	-
2440	1	`yüzde9açıkar`	0	1	-	0,1,0
2460	1	`yüzü` (Gesicht|POSS)	1	0	2	-
2480	1	`yortabiibu`	0	1	-	0,1,0
2500	1	`ıyorancak`	0	1	-	0,1,0
2520	1	`yokgibi`	0	1	-	0,1,0
2540	1	`şöylesıral`	0	1	-	0,1,0
		sum	1192 78.7%	327 21.5%		318 (20.9%), 7 (0.5%), 1 (0.1%)

Tabelle 2.4: Türkische Beispielsegmente und ihr linguistischer Status. *Rang*: Platz des jeweiligen *Segmentes* in der sortierten Frequenzliste aller *Segmente* an. *Häuf*(igkeit): Zahl der Vorkommen im Output. *korrekt*/*falsch*: meine Beurteilung der Vorkommen. Sie summieren sich nicht immer zur Gesamthäufigkeit auf, da aus technischen Gründen nicht wirklich alle *Segmente* beurteilt wurden. *min. Seg.*: Zahl der *minimalen Segmente* (≈Morpheme), die in der Zeichenkette maximal enthalten sind. *Fehler*: Fehler in der Notation 2a,2b,2c auf.

Eine weitere sehr wichtige Beobachtung ist die Tatsache, dass die überwältigende Mehrzahl aller Fehler in die Kategorie 2a fällt, also reine Übersegmentierungen darstellt. Fehler der Kategorien 2b und 2c fallen mit $< 1\%$ kaum ins Gewicht. Es ist aber wichtig im Auge zu behalten, dass diese Ergebnisse sich auf einen Querschnitt der Daten beziehen und auf die Gesamtmenge der Segmente nur schwer zu verallgemeinern sind. Die Mehrzahl der Segmente in den Frequenzlisten der Segmente sind *Hapax Legomena*.[82] In den gezeigten Tabellen sind sie unterrepräsentiert.

Es fällt auf, dass ein großer Teil der Segmente, die in die Fehlerkategorie 2b fallen, zwar für sich alleine genommen keine Bedeutung haben, aber durch eine Erweiterung links oder rechts zu einer eigenständigen Bedeutung gelangen.

Willkürlich ausgewählte Beispiele aus allen drei Sprachen sind:

Deutsch `anderspitzeder`, `darindaß`, `dasdeutsche`, `diegefahrdes`, `tezujenerzeit` (mit der Präteritumsendung *te*)

Englisch `accordingto`, `describedas`, `helookedso`, `matterof`, `outofthe`

Türkisch `dentamam/dantamam` („die Zustimmung von X“), `egöre/agöre` („gemäß X“), `büyükbir` („ein großer X“), `labirlikte` („zusammen mit X“), `herhangibir` („irgendein X“)

Derartige Strukturen lassen einen unwillkürlich an Begriffe wie die von Biber eingeführten *Lexical Bundles* (Biber, 1999; Biber und Barbieri, 2007) denken. Im gegenwärtigen Kontext liegt der Vergleich mit *Lexical Bundles* durchaus nahe, da er wie das hier vorgestellte Segmentierungsverfahren rein häufigkeitsbasiert gedacht werden kann. Biber führt allerdings eine feste Längenbeschränkung ein und arbeitet mit tokenisiertem Text. Diese Beschränkungen fallen hier fort.

Die Stellung der angesprochenen Leerstelle zeigt Unterschiede: Im Deutschen und im Englischen steht sie rechts, einzige Ausnahme ist das deutsche `tezujenerzeit`. Im Türkischen dagegen ist sie mehrheitlich links, bis auf `herhangibir` („irgendein X“) und `büyükbir` („ein großer X“), die eine rechte Erweiterung fordern. Diese Beobachtung ist zwar durch fünf Beispiele pro Sprache kaum zu untermauern, korrespondiert aber mit der Tatsache, dass Türkisch eine SOV-Sprache ist, im Gegensatz zu den SVO-Sprachen Deutsch und Englisch. Ein weiterer Stellungsunterschied ist, dass das Türkische nur Postpositionen kennt, während in den beiden anderen Sprachen Präpositionen wesentlich häufiger sind.

In welchem Kontext finden sich die auftretenden Fehler? Betrachten wir die 15 falschen *Segmente* `en`. Diese verteilen sich folgendermaßen:

- 4 befinden sich direkt vor oder nach einem unsegmentierbaren Textstück, also nach einer sogenannten *Brücke* (s. 2.5.2). Es ist nachvollziehbar, das hier besondere Schwierigkeiten auftreten.
- 3 sind innerhalb eines Namens oder Fremdwortes. Auch hier ist mit Fehlern zu rechnen.

[82] 75% für Deutsch, 80% für Englisch und 74% für Türkisch

- 3 sind eine Übersegmentierung des bestimmten Artikels `d|en`. Hier könnte man sogar auf dem Standpunkt stehen, dass die Segmentierung nicht ganz ohne Sinn ist. Die bestimmten Artikel und Demonstrativpronomina stellen eine Serie dar, die durchweg mit einem `d` beginnt (*der, die, das, dieser, diese, dieses...*). Dieses `d` kann unter Umständen als *minimales sprachliches Segment* interpretiert werden. Diese diachrone Sichtweise ist synchron zwar kaum zu rechtfertigen. Im Goldstandard kommt eine Zerlegung von `dem` allerdings tatsächlich vor, wenn auch nicht in `d em`, sondern in `de m`.

- Die verbleibenden 4 Fehler sind weniger einheitlich. Ein interessantes Beispiel ist aber `ein|en|acht`, wo es richtig heißen müsste `eine|nacht`. Bemerkenswert ist hier, dass es sich hier um eine Aneinanderreihung von möglichen Morphemen handelt. Nur ergeben sie im gegebenen Kontext keinen Sinn. Dies ist ein häufiger Fall. Einschränkend ist zu erwähnen, dass ein großer Teil derartiger Mehrdeutigkeiten erst entstehen, wenn die Leerzeichen aus dem Text entfernt werden. In der Zeichenkette `eine_nacht` ist der Substring `e_n` nicht mehr mit dem deutschen Flexisonssuffix `en` zu verwechseln.

Ein interessanter Fall ist auch `ich`, das bei 69 beurteilten Vorkommen 11 Fehler aufweist. `ich` ist im geschriebenen Deutschen eindeutig genau einem Morphem zuzuordnen, es kommt nur als Personalpronomen 1. Person Singular vor. Alle 11 Fehler sind darauf zurückzuführen, dass es eine Tendenz gibt, dieses sehr häufige Element auch dann zu erkennen, wenn es in einem bestimmten Kontext nicht möglich ist. Betrachtet man die 266 Vorkommen der Zeichenkette `ich` im Testkorpus genauer, findet man darunter 62 Personalpronomen. Dies entspricht einem *Recall* von 0.91, einer *Precision* von 0.86 (und einem f von 0.89).[83] In Anbetracht der Tatsache, dass das System über keinerlei grammatisches Wissen verfügt, ist es sehr positiv zu werten, dass unter den 266 Vorkommen er Zeichenkette `ich` nur so wenige falsch positive auftreten.

Man kann festhalten: Der Grundgedanke der Definition von *Segmentierungen* (Definition 19) als lückenlose Ketten möglicher Segmente trägt Früchte, da ein sehr großer Anteil potentiell mehrdeutiger Zeichenketten durch die Berücksichtigung des Kontextes korrekt segmentiert wird. Dies führt zu einer relativ hohen *Precision*. Beruhigend ist auch der kleine Anteil der als besonders schwerwiegend einzustufenden Fehler der Kategorie 2c.

Dennoch bleibt eine Fehlerrate von 15 bis 20% bestehen. Dieser Wert kann aus den Tabellen 2.2 bis 2.4 abgelesen werden. Dieselbe Größenordnung findet sich für *representation* = `s` auch in Abbildung 2.11, wo es um den Anteil der korrekt erkannten Leerzeichen geht. Creutz und Lagus (2002), eine Arbeit mit ähnlicher Zielsetzung, – s. auch Seite 40 – geben sehr ähnliche Fehlerquoten an, obwohl die Ergebnisse nicht vollständig vergleichbar sind. So widmen sich Creutz und Lagus der Zerlegung von finnischen Wörtern, während es hier um die Zerlegung leerzeichenbereinigter Sätze drei anderer Sprachen geht.

Insgesamt scheinen die hier erzielten Ergebnisse auf den ersten Blick mindestens

[83] Ganz nebenbei wird damit die Annahme, dass *Precision* und vor allem *Recall* für alle Segmente ungefähr gleich sind, erschüttert.

konkurrenzfähig zu sein. Relativ zu anderen automatischen Verfahren sind sie das wohl auch.

Aber: Für einen deutschen, englischen oder türkischen Muttersprachler ist es zwar mühsam, einen Text ohne Leerzeichen zu lesen, er wird aber im Allgemeinen alle Wörter korrekt erkennen und segmentieren können.[84] Die 15% Performanzunterschied zwischen Mensch und Maschine sind natürlich darauf zurückzuführen, dass der Mensch den Text versteht, während der Computer nur Häufigkeiten zählt. Der menschliche Leser weiß, dass die Segmentierung `myhusband|is|ing|re|at|pain`[85], falsch sein muss, unter anderem weil nach der Verbform `is` kein `ing` folgen kann. Für den Algorithmus dagegen ist kein Problem erkennbar, er hat die Zeichenkette in Segmente zerlegt, die unter anderen Umständen alle englische Morpheme repräsentieren können. Die Probleme folgen also daraus, dass der Algorithmus keine Information über linguistische Kategorien hat und dieses auch nicht zu lernen versucht. Ähnliche Schwierigkeiten zeigen sich nicht nur hier, sondern in anderen bisher veröffentlichten Arbeiten zur automatischen Morphologieanalyse. Die Hartnäckigkeit und die prinzipielle Natur des Problems wird auch von Goldsmith (2010) erkannt. Er bezieht sich auf das verwandte Feld der *Wort*segmentierung, dh. auf die Zerlegung von unsegmentiertem Text in Einzelwörter. Seine Erkenntnisse und Schlussfolgerungen sind aber ohne große Änderungen auch auf die Morpheminduktion zu übertragen:

> The most interesting result of all of the work in this area is this: there is no way to solve the word segmentation problem without also making major progress with the problem of automatic learning of morphology and syntax. Knowledge of the statistical properties of strings can be used to infer words only to the extent that the device that generated the strings in the first place *used* knowledge of words, and only knowledge of words, to generate the string in the first place; and, in actual fact, the systems that generate our natural language strings employ systems at several levels: it is not words, but *morphemes* that consist of relatively arbitrary sequences of letters, and words are the result of a system responsible for the linear placement of morphemes. In addition, there is a system responsible for the sequen- tial placement of words – we call it *syntax* – and it too has a great impact on the statistics of letter placement. A system that tries to learn the structure of language on the basis of a model that is far poorer than the real structure of language will necessarily fail – we may be impressed by how well it does at first, but failure is inevitable, unless and until we endow the learning algorithm with the freedom of thought to consider models that take into consideration the structure that indeed lies behind and within language. (Goldsmith, 2010, S. 380f, Hervorhebungen im Original)

[84] Es gibt Ausnahmen. Die Zeichenkette

`nachdem ende oktober letzten jahres in frankfurtamlsonnemann gestorben ist`

erwies sich als sehr schwer verständlich. Leerzeichen und Fettdruck hier zur Verdeutlichung.

[85] So in den Ergebnissen zu finden.

Einige Autoren versuchen, mit dem Problem umzugehen. So führt bereits Goldsmith (2001) verschiedene Kategorien von Segmenten ein. Sein Algorithmus kennt aber nur *Stamm* und *Suffix*. Creutz und Lagus gehen in ihren verschiedenen Arbeiten einen ähnlichen Weg. Die eingeführten grammatikähnlichen Systeme haben aber immer einen starken *ad-hoc*-Charakter, bzw. implementieren sprachliches Wissen.

Die erwähnten Probleme sind auch die Motivation hinter dem von Schone und Jurafsky (2000, 2001) vorgeschlagenen Algorithmus (vgl. Abschnitt 2.4.1 auf Seite 36). Schone und Jurafsky bemühen sich, ebenfalls auf statistischem Weg, semantische Ähnlichkeiten abzuschätzen und diese zur Identifizierung falscher Segmentierungen zu nutzen. Eine wirklich tragfähige Lösung, die der menschlichen Performanz auch nur entfernt nahe kommt, steht aber noch aus.

Wie könnte ein möglicher Ausweg aussehen? Ein in diesem Zusammenhang interessanter Fall ist im türkischen Testtext das Segment `yor`.[86] Es ist nur selten ein eigenständiges *sprachliches Segment* des Türkischen.[87] Die Tatsache, dass der Algorithmus es als solches vorschlägt, ist auf eine Besonderheit der Vokalharmonie an dieser Stelle zurückzuführen. Im Allgemeinen treten türkische Endungen in mehreren Formen auf, je nachdem, welchen Vokal die davor liegende Silbe enthält. So existiert das Pluralmorphem als `ler` und `lar`, bzw die Verbindung `iz`, die die 1. Person Plural repräsentiert, erscheint je nach Kontext als `iz`, `ız`, `uz` oder `üz`. Das Präsens anzeigende `iyor` ist eine Ausnahme, da nur der erste Vokal den Harmonieregeln unterliegt, das für eine türkische Endung ungewöhnliche `o` bleibt unverändert. Damit gibt es die vier Varianten `iyor`, `ıyor`, `uyor` und `üyor`. Vor der Zeichenkette `or` ist in der Hälfte der Fälle ein `y` zu finden. In 97% der Fälle ist das entstehende `yor` Teil der erwähnten Endung. Die häufigste Variante `iyor` hat mit 44% Häufigkeit `Predictability change` kleiner als 1. Daher wird die Endung sehr häufig übersegmentiert und das falsche Segment `yor` entsteht.

Dieses Beispiel wurde deshalb so ausführlich dargestellt, weil sich in diesem speziellen Fall ein allgemeiner Lösungsansatz besonders deutlich abzeichnet. Die Zeichenkette `yor` ist nicht der einzige Kontext, den die Vokale `i`, `ı`, `u` und `ü` teilen. Entsprechend der Vokalharmonie zeigen eine ganze Reihe Endungen wie das erwähnte `iz`, aber auch `miş`, oder `siniz` sehr ähnliche Muster: Sie kommen in allen vier Varianten vor, jeweils mit einem ähnlichen Frequenzverhältnis. Da dieses Muster ausgesprochen regelmäßig und in jedem türkischen Text sehr häufig ist, sollte es möglich sein, es automatisch zu extrahieren. Das Ergebnis eines solchen hypothetischen Algorithmus wäre die Feststellung, dass die vier Vokale `i,ı,u,ü` zu einer Kategorie vereinigt werden können. Unter der Annahme, dass die vier Vokale als Kategorie, also als eine einzige Entität begriffen werden, sollte das falsche Segment `yor` verschwinden und der korrekten linguistischen Einheit `{iıuü}yor` Platz machen. Ein ähnlich fruchtbares Beispiel wäre die mehrfach erwähnte Groß- und Kleinschreibung im Deutschen. Es wäre wahrscheinlich ein großer Schritt vorwärts, wenn *haus* in `das Haus` und `Wohnhaus` als zusammengehörig erkannt würde ohne dabei Information zu verlieren.

[86] In Tabelle 2.4 kommt es nicht vor. Es steht mit 71 Vorkommen an 16. Stelle der Frequenzliste.

[87] Es gibt das Verb *yor|mak*, es ist nicht besonders häufig und fällt gegenüber der der Präsensendung *iyor* nicht ins Gewicht.

Damit wäre nicht nur eine Detailfrage besser gelöst, sondern der Grundgedanke wäre qualitativ ein völlig anderer. Es ginge nicht mehr nur darum, aus vorhandenen Frequenzinformationen eine Segmentierung des Textes zu errechnen, sondern gleichermaßen darum, die entstehenden Einheiten kontextabhängig in Kategorien einzuordnen. Wenn es gelänge, beide Teile rekursiv miteinander zu verbinden, so dass jede entdeckte Kategorie die Segmentierung in linguistische Einheiten verbessert und umgekehrt, so wäre der Weg frei zu umfassender Grammatikinferenz.

2.7 Zusammenfassung und Diskussion

Es folgt eine Zusammenfassung der Untersuchungen zur *Morphologischen Induktion* und ihrer Ergebnisse. Das Kapitel beginnt mit einer Klärung der notwendigen morphologischen Grundbegriffe (Abschnitt 2.2). Da diese in der aktuellen Forschung eine gewisse Unschärfe aufweisen, ist es notwendig, darzustellen, was in dieser Arbeit mit welchem Terminus bezeichnet wird. An zentraler Stelle steht die Definition des *(minimalen) sprachlichen Segments* (Definitionen 6 und 9). So bezeichne ich Zeichenketten, die Elemente der Formseite sprachlicher Zeichen sind. Diese Formseiten sind als Mengen definiert um diskontinuierliche Zeichen zuzulassen. Der Algorithmus wird daran gemessen wie gut er in der Lage ist, die *sprachlichen Zeichen* des Textes zu dekodieren.

In Abschnitt 2.4 werden wesentliche Strömungen der bisherigen Forschung zur *Morphologischen Induktion* referiert. Zwei Grundansätze lassen sich unterscheiden: Auf der einen Seite stehen heuristische Methoden, die direkt an den von Harris (1955) erstmals explizit aufgestellten Gedanken anknüpfen, dass linguistische Einheiten aus Elementen bestehen, die oft zusammen vorkommen, untereinander aber frei kombiniert werden können. Auf der anderen Seite stehen Ansätze, die eine Klasse von Sprachmodellen aufstellen und mit Hilfe eines Suchalgorithmus dasjenige auswählen, dass am besten zum Ausgangstext passt. Dies sind die sogenannten Bayes'schen Ansätze.

Der hier vorgestellte Algorithmus ist eine radikale Interpretation der ersten Idee. Er arbeitet auf Daten, die noch nicht in dieser Vollständigkeit zu diesem Zweck eingesetzt wurden. Dabei verfolgt er eine sehr weitgehende Zielsetzung: Es sollen nicht nur *minimale sprachliche Segmente* gefunden werden, sondern diese auch zu größeren Strukturen zusammengeordnet werden.

Der Grundaufbau des Algorithmus ist zweiteilig. Der erste Teil ist eine Formalisierung des Gedankens, dass die Vorhersagbarkeit des jeweils nächsten Zeichens an beiden Seiten eines *Segments* abfällt. Ergänzend dazu wird gefordert, dass der Text aus einer lückenlosen und überschneidungsfreien[88] Zerlegung in Segmente bestehen muss.

Die verbleibende Mehrdeutigkeit wird durch ein Ranking-Verfahren aufgelöst. Vier kategoriale Parameter legen die Strategie des Systems auf verschiedenen Ebenen fest. Der lokale Parameter P_L bestimmt, wie die einzelnen *Segmente* in die Bewertung eingehen. Der Parameter P_T gibt vor, mit welchem Güteindex Bäume aus Segmenten und Teilsegmenten belegt werden. P_4 entscheidet über die Bewertung von Kindsegmentpaaren, falls

[88] Nur Leerzeichen dürfen zu aufeinanderfolgenden *Segmenten* gehören.

es hier mehrere Möglichkeiten gibt. P_F regelt wie die Bewertungen aufeinanderfolgender Segmentbäume zur Gesamtbewertung einer Segmentierung verbunden werden.

Bei der Evaluation des Algorithmus stellt sich folgendes Problem: Ein allgemeingültiger Goldstandard ist nicht vorhanden und kann aufgrund der erwähnten theoretischen Unschärfe auch nicht existieren. Dem Problem wird mit drei sich gegenseitig ergänzenden Auswertungsverfahren Rechnung getragen. In einem ersten Schritt werden nur die Leerzeichen überprüft, bzw. die Stellen, an denen im Originaltext Leerzeichen standen. Damit steht unmittelbar eine große Menge unstrittiger Segmentgrenzen zur Verfügung. Nachteilig ist, dass diese Untermenge an Segmentgrenzen nicht zufällig ausgewählt ist: Es scheint plausibel, dass wortinterne Grenzen schwieriger zu finden sind als Grenzen zwischen orthographischen Wörtern, unabhängig davon, ob die Leerzeichen entfernt wurden. Aus diesem Grund könnte es zu optimistisch sein, die aus den Wortgrenzen gewonnenen Ergebnisse zu verallgemeinern. Daher wird in einem zweiten Schritt doch ein, wenn auch kleiner, Goldstandard für einen Teil des deutschen Korpus erstellt. Der unumgänglichen theoretischen Unschärfe begegne ich durch die unabhängige Befragung dreier Experten. Da es nun nicht mehr nur rein kategoriale Entscheidungen für oder gegen Segmentgrenzen gibt, müssen die Evaluationsmaße *Recall*, *Precision* und *F-Measure* angepasst werden. In einem abschließenden dritten Evaluationsschritt wird eine Auswahl an Segmenten aus jeder Sprache gesondert betrachtet und beurteilt.

Die vier Parameter P_L, P_T, P_F und P_4 spannen eine Vielzahl an möglichen Kombinationen auf und wechselwirken stark miteinander. Um ihre Wirkung dennoch übersichtlich quantifizieren zu können wird die Performanz des Algorithmus in Abhängigkeit der Parameter modelliert. Dabei kommen *lineare gemischte Modelle* zum Einsatz. Für die Erkennung der Leerzeichen wird die *Performanz* des Algorithmus mit einem binomial verteilten Fehlerterm modelliert. Die *Performanz* kann hier als eine eingeschränkte Variante des *Recalls* verstanden werden. In Bezug auf den kleinen deutschen Goldstandard können *Recall*, *Precision* und F-Measure[89] getrennt ausgewertet werden. Hier werden jeweils normalverteilte Fehlerterme angenommen.

Die in diesem Ansatz enthaltenen Näherungen werden ausführlich diskutiert. Im Ergebnis kann angenommen werden, dass die Unzulänglichkeiten der Modelle in Bezug auf die Daten zwar einen Einfluss auf die numerischen Werte der p-Werte haben können, sich aber nicht auf die relativen Verhältnisse der Parameterschätzungen auswirken dürften und die Vergleichbarkeit der drei untersuchten Sprachen kaum beeinflussen sollten. Man gewinnt mit diesem Vorgehen eine erhebliche Auflösung. So waren selbst bei Betrachtung der lediglich 20 Sätze des Goldstandards feinste Wechselwirkungen zu erkennen. Darüber hinaus zahlt es sich aus, dass in den beiden empirischen Untersuchungen unterschiedliche aber vergleichbare Daten, Modelle und Softwarealternativen für die Evaluation verwendet werden. Die so gewonnenen Resultate können sich gegenseitig stützen und ergänzen.

Es ergibt sich eine Reihe berichtenswerter Ergebnisse, sowohl in Bezug auf mögliche Anwendungen, als auch aus theoretischer Hinsicht. Mit Blick auf zukünftige Anwendungen sollen hier zwei Eigenschaften hervorgehoben werden.

[89] bzw ihre angepassten Versionen.

Zum Einen ist die allgemeine Performanz der Methode hoch: Obwohl, wie sorgfältig dargelegt wird, eine wirkliche quantitative Vergleichbarkeit bei der Struktur der derzeitigen Forschungslandschaft unmöglich bleibt, kann der Algorithmus in Bezug auf sein Kernziel (die Zerlegung von Text in *minimale sprachliche Segmente*) als höchst wettbewerbsfähig gelten. So ist in Abbildung 2.10 und 2.11 zu erkennen, dass der Median für den Anteil der als *Segmentgrenzen* erkannten Leerzeichen im deutschen Text in Originalform bei etwa 0.98 liegt. Im Englischen liegt er leicht darunter und auch im Türkischen gibt es noch eine erhebliche Zahl an Sätzen, in denen alle Leerzeichen erkannt werden.

Zum zweiten ist es für zukünftige Anwendungen eine sehr wichtige Tatsache, dass der Parametersatz mit optimaler *Performanz* sich in allen drei Sprachen als identisch herausstellt. Schreibt man für eine bisher ungesehene Sprache die Parameter auf diese optimalen Werte fest, ergibt sich ein vollständig unüberwachter und sprachunabhängiger Algorithmus, da kein weiteres sprachliches Wissen eingeht. Es ist sehr wohl möglich, dass sich diese sehr günstige Eigenschaft nicht wirklich auf alle Sprachen übertragen lässt. Hier ist noch Raum für weitere Forschung. Aber gerade in der Nähe der optimalen Werte gibt es nur sehr geringe *Performanz*-Unterschiede, so dass sich kein großer Verlust ergeben sollte, so lange P_L den Wert `combined` annimmt.

Die stabile Überlegenheit dieses Parameters ist ein weiteres wesentliches und auch theoretisch interessantes Ergebnis der Untersuchung. Die mit $P_L =$ `combined` verbundene Ranking-Strategie kombiniert die Vorhersagbarkeitsabfälle an beiden Seiten eines *möglichen Segmentes* zu einer logarithmischen Summe: Die logarithmische Transformation übersetzt Produkte in Summen, aufgrund der allgemeinen Beziehung $\log(ab) = \log(a) + \log(b)$. Nun sind die Vorhersagbarkeitsabfälle $D^{\pm}$ Produkte und Brüche aus Häufigkeiten. Infolgedessen übersetzt der Logarithmus sie in Summen von Logarithmen von Häufigkeiten. Es ist ein Ergebnis der berichteten Untersuchungen, dass es für die Präzision des Segmentierungsalgorithmus von Vorteil ist, die eingehenden Substringhäufigkeiten auf der logarithmischen Ebene zu vergleichen, im Gegensatz zu den absoluten Zählungen. Dies gibt den kleinen Frequenzen mehr Gewicht.

Es lässt sich ebenso als allgemeines Charakteristikum festhalten, dass Segmentierungen mit möglichst vielen Elementen sich tendenziell als die besseren herausstellen. Diese Beobachtung korrespondiert mit der Tatsache, dass die *(weighted) Precision* in Bezug auf den deutschen Goldstandard über dem *(weighted) Recall* liegt. Das heißt, es werden nicht alle Grenzen *sprachlicher Segmente* gefunden, aber die vorgeschlagenen sind oft korrekt. Es kann in diesem Zusammenhang durchaus diskutiert werden, ob bzw. in Bezug auf welche Zielsetzung ein maximaler *Recall* als optimal anzusehen ist. Verbindungen mehrerer *sprachlicher Segmente* manchmal unanalysiert zu lassen, erscheint zumindest mit Blick auf die Repräsentation von Sprache im menschlichen Gehirn nicht unangemessen.

Ein weiteres Phänomen scheint an zwei unterschiedlichen Stellen der Evaluation auf: Für die *Performanz* der Rückgewinnung der Leerzeichen ist die Strategie $P_L =$ `forward`, die die *Segmente* nach ihrem *Forward Predictability Change* bewertet ihrem Gegenstück $P_L =$ `backward` überlegen. Bei P_4, dem Parameter, der das Ranking verschiedener Kindsegmentpaare steuert, ergibt sich ein gegenläufiger Effekt. Hier ist es günstiger, die Elemente nach ihrem *Backward Predictability Change* zu bewerten. Dass es überhaupt

Richtungseffekte gibt, ist angesichts der eindeutigen Richtung, die jeder Sprachäußerung innewohnt, keine Überraschung. Nach meinem Kenntnisstand ist aber die vorliegende Untersuchung der erste empirische Nachweis eines solchen Effektes auf dem Gebiet der *Morphologischen Induktion*. Und auch wenn die Existenz einer Vorwärts-Rückwärts-Asymmetrie in natürlichsprachigen Texten als gegeben angenommen werden kann, so könnte es sich doch als fruchtbar erweisen, ihre genaue Struktur in verschiedenen Sprachen und auf verschiedenen Ebenen erst empirisch zu erfassen und dann durch Modelle zu erklären zu suchen.

Es ergeben sich zahlreiche Denkansätze für weitere Forschung. Dazu gehört zum einen die weitere systematische Untersuchung sorgfältig ausgesuchter und neuer Varianten der Parameter $P_{L,T,F,4}$ in verschiedenen Sprachen. Ebenso wären Untersuchungen mit Hilfe weiterer Goldstandards in weiteren Sprachen wünschenswert. Auch zur Überlegenheit des Logarithmus und den beobachteten Asymmetrien wären weitere Details aus vergleichenden Untersuchungen wertvoll.

Der vielleicht faszinierendste Gedanke liegt aber in der Frage, ob es möglich ist, den bisherigen Algorithmus um einen zweiten, komplementären Teil zu ergänzen. Dieser würde zwei Beobachtungen in Rechnung stellen: Auf höherer Ebene werden tendenziell keine vollen sprachlichen Segmente, sondern Schablonen für ganze Klassen von Segmenten erkannt. Auf unterster Ebene entstehen viele Fehler daraus, dass das System keinerlei Wissen darüber hat, welche Segmente wie miteinander kombiniert werden können. Es ist denkbar, dass sich Kategorien lernen lassen auf eine Art und Weise, die beide Probleme gleichermaßen angeht.

3 Stilometrie

3.1 Einleitung

Thema der vorliegenden Arbeit sind vollständige Substringfrequenzen natürlichsprachiger Texte. Diese Daten wurden bisher keiner systematischen Analyse unterzogen, obwohl verschiedene Indizien ihr Potential andeuten. Ich stelle zwei Anwendungen auf Grundlage dieser Daten vor. Die Zielsetzung ist nicht nur die Demonstration der praktischen Nutzbarkeit vollständiger Substringfrequenzen. Darüber hinaus lassen sich aus den Details der Performanz der Algorithmen vielfältige und fruchtbare empirische Folgerungen ableiten. Diese bergen Relevanz auch für das Verständnis der Rolle von Häufigkeiten im System der Sprache in sich und sind damit linguistisch potentiell von großer Bedeutung.

Im vorigen Kapitel wurde ein Verfahren zur automatischen, sprachunabhängigen und (letztendlich) parameterfreien Erkennung von Morphologie in untokenisiertem Text vorgestellt. Aufgrund der Struktur des Problems lag bei dieser Untersuchung der Fokus auf den lokalen Häufigkeitsverhältnissen. Dies äußert sich im zentralen Begriff der *predictability*, bzw. des *predictability change*. Das heißt, obwohl der gesamte Kontext innerhalb des zu segmentierenden Abschnitts in die Segmentierung einfließt und obwohl die globalen Häufigkeitsverhältnisse im *Trainingskorpus* in die Analyse einbezogen werden, ist die Segmentierung in Bezug auf den *Testtext* unweigerlich ein wesentlich lokaler Prozess.

Im folgenden Kapitel wird nun mit der *Stilometrie* ein Bereich behandelt, in dem das Gewicht auf dem globalen Vergleich von Texten liegt und damit auf den Eigenschaften der Susbstringfrequenzstatistiken des jeweiligen Textes als Ganzes.

Was genau ist *Stilometrie*? Dem Namen nach misst Stilometrie Stil. Auf die Definitionen und Diskussion des Begriffes „Stil“ wird allerdings in der mir bekannten Forschungsliteratur kein Bezug genommen. Einer wenn auch informellen Definition des Begriffs *Stilometrie* jedoch nähern sich Clement und Sharp (2003), wenn sie schreiben: „It could be said that any author, amateur or professional, whose documents show elements of consistency from one to another can be considered to have elements of ‚style‘.“[1] Der Grund für das Fehlen einer formalen Definition für *Stil* in der *Stilometrie* scheint im Zitat deutlich durch: Es wird zwar angenommen, dass es messbare Eigenschaften der persönlichen Sprache gibt, auf die wir höchstens eingeschränkten bewussten Zugriff haben. Bei genauerem Hinsehen geht es bei den als *stilometrisch* bezeichneten Ansätzen

[1] Eine ähnlich formulierte Definition von *Stilometrie* findet sich bei Juola (2006a): „We can thus define ‚authorship attribution‘ broadly as any attempt to infer the characteristics of the createor of a piece of linguistic data.“ (Juola verwendet den Begriff *authorship attribution* als „near-synonymous“ mit *Stylometrie*.)

aber nicht um eine qualitative oder quantitative Bestimmung dieses *Stils* an sich, sondern um die Klassifikation[2] von Texten nach einer bestimmten Klasse von Variablen. Am Anfang stand die Frage nach der Autorenschaft eines Textes. So definiert Holmes (1998) Stilometrie noch als „the statistical analysis of literary style“ und schreibt etwas später im selben Text: „[...] stylometrists hope to uncover the ‚characteristics‘ of an author“. Im letzten Jahrzehnt aber weitet sich der Blick, wie ein Zitat aus Gamon (2004) zeigt: „The identification of authorship falls into the categorization of style classification, an interesting sub-field of text categorization that deals with properties of the form of linguistic expression as opposed to the content of a text“. Heute spielen neben der Autorenschaft auch andere Klassifikationskriterien eine Rolle. Beispiele umfassen Klassifikation nach dem Geschlecht des Autors (z.B. Argamon et al. (2003)), seiner Muttersprache (z.B. Zigdon (2005); Koppel et al. (2005); Tsur und Rappoport (2007)) oder in übersetzte und nicht übersetzte Texte (*Translationese*, z.B. Baroni und Bernardini (2006); van Halteren (2008); Ilisei et al. (2010)). Derartige Variablen bezeichne ich als *stilometrische Variablen*. Ihnen ist gemein, dass sie text-, bzw. sprachexterne Eigenschaften messen, im Gegensatz zu Variablen wie Sprache, Dialekt, Topic oder Genre[3]. Tweedie et al. (1996) gehen nur scheinbar über eine rein klassifikatorische Sicht der *Stilometrie* hinaus, wenn sie schreiben: „We define style as a set of measurable patterns which may be unique to an author“. Auch ihre Arbeit beschäftigt sich mit Klassifikation und nicht mit der Messung von Stil an sich. Diese Beschränkung überwinden erst Koppel et al. (2009) bis zu einem gewissen Grad, mit ihrer Zielsetzung der *Authorship Verification*. Ich gehe in Abschnitt 3.6.4 durch den Vergleich der Ähnlichkeit der Texte von ein- und zweieiigen Zwillingen über die reine Klassifikation hinaus.

Zusammenfassend gesagt, versucht die *Stilometrie* Ähnlichkeiten in den Häufigkeitsverteilungen von oberflächennahen Texteigenschaften zu quantifizieren und zur Klassifikation dieser Texte nach einer bestimmten Klasse von Variablen zu nutzen.

Stilometrie ist durchaus nicht nur als Grundlagenforschung von Bedeutung. Das trifft vielleicht in besonderem Maße auf die Automatische Autorenbestimmung (AA) zu, die natürlich ein Hilfsmittel für (Literatur-)Historiker darstellt, oder im Rahmen der so genannten forensischen Linguistik vor Gericht Expertenwissen einbringen kann (Chaski, 2001, 2005). Man kann sogar vermuten, dass stilometrische Methoden im nachrichtendienstlichen Bereich eine gesamtgesellschaftliche wenn auch verdeckte Relevanz bekommen (Estival et al., 2008).

Die folgenden Untersuchungen beschäftigen sich mit *Stilometrie* aber tendenziell um ihrer selbst willen, bzw. aus grundlegendem wissenschaftlichem Interesse an den allgemeinen Schlussfolgerungen, die sich aus den Ergebnissen ziehen lassen. Zwei Hauptfragen gilt besondere Aufmerksamkeit. Die erste motiviert sich aus einem hauptsächlichen Ergebnis der Evaluation des Segmentierungsalgorithmus in Kapitel 2. Dort ergibt sich eine klare Überlegenheit der logarithmisch transformierten und aufsummierten Daten

[2] „style-based classification“ (Stamatatos, 2009, 540)

[3] Der Begriff *Topic* hat in der Stilometrie eine andere und wesentlich einfachere Bedeutung als auf anderen Gebieten der Linguistik. In der Stilometrie ist dieser Terminus lediglich eine Bezeichnung für den Inhalt oder das Thema eines Textes. Ähnlich allgemein bezeichnet *Genre* die Art oder Sorte des Textes. Für *Topic* und *Genre* finden in der Stilometrie ebenso wenig wie für *Stil* feste Definitionen.

(P_L = `combined`) gegenüber der direkten Verwendung der absoluten Häufigkeitszählungen. Hier bietet sich nun die Möglichkeit, die Bedeutsamkeit derselben Transformation derselben Daten in einem unterschiedlichen Kontext zu untersuchen. Dies eröffnet empirische Möglichkeiten, kategorische Aussagen wie die folgende kritisch zu hinterfragen: „Note that [...] the most frequent character n-grams are the most important features for stylistic purposes“ (Stamatatos, 2009, 541). Das Gewicht lag in den letzten Jahrzehnten mehr und mehr so stark auf den häufigen Textelementen, dass die Untersuchung seltenerer Bestandteile vernachlässigt wurde.[4]

Die zweite Fragestellung dagegen ist spezifisch für die *Stilometrie*. *Morphologische Induktion* ergibt nur in Bezug auf den rohen, unannotierten (allenfalls tokenisierten) Text Sinn, da eine Zerlegung des Textes in linguistisch relevante Einheiten der erste Schritt jeder Analyse sein muss. Im Gegensatz dazu lassen sich stilometrische Methoden im Allgemeinen auf verschiedene Annotationsebenen eines Korpus anwenden, so zum Beispiel auf die POS-Tag-Sequenz der Texte oder ihre Lemmatisierungen. Durch den Vergleich derselben stilometrischen Methodik auf den verschiedenen Annotationsebenen ist es möglich, sich der Frage zu nähern, auf welcher Ebene welche Art von Information angesiedelt ist.

Es ist an sich keine neue Idee, *Zeichen-n*-Gramme im Rahmen stilometrischer Verfahren einzusetzen (Details im folgenden Abschnitt), alleine schon deshalb, da für den Computer nutzbare Daten entweder auf oberflächennahe Eigenschaften beschränkt sind, oder eine aufwendige Annotation erfordern. Da die reine Zeichenkette des Textes die am einfachsten zugängliche Datenquelle ist, liegt es nahe, hier mit der Analyse zu beginnen. Meines Wissens allerdings wurden die *vollständigen Substringhäufigkeiten*, die das Thema meiner Arbeit darstellen, noch nie direkt für stilometrische Zwecke verwendet. Die einzige mir bekannte Arbeit, die auch Suffixbäume in diesem Kontext verwendet (und damit potenziell dieselben Daten untersucht), benutzt sie vor allem zur Filterung und Extraktion besonders „relevanter“ n-Gramme (Zhang und Lee, 2006).

Es gilt nun, die vollständigen Substringhäufigkeiten in einen effektiven stilometrischen Algorithmus umzusetzen. In einem ersten Schritt definiere ich ein Maß, das die Ähnlichkeit zweier solcher Verteilungen quantifiziert. Aufgrund der so berechneten Ähnlichkeit werden die Texte klassifiziert.

Es sind viele Ähnlichkeitsmaße vorstellbar, die sich auf derartigen Frequenzdaten erklären lassen. Ich werde eine Familie von Maßen vorstellen, die als Funktionen zweier Frequenzvektoren aufgefasst werden können. Vor allem die erfolgreicheren Varianten unterscheiden sich substanziell von den in der bisherigen stilometrischen Forschung verwendeten vergleichbaren Textähnlichkeitsmaßen.

Die vorgestellte Methode in ihren Varianten wird anhand verschiedenartiger Fragestellungen evaluiert: Automatische Autorenbestimmung (Abschnitte 3.5 und 3.6.3), der Vergleich übersetzter und originaler Texte (*Translationese*, Abschnitt 3.6.1), die Klassi-

[4] „With regard to the choice of features, there is a growing consensus that analysis of high frequency words (mostly function, or closed class, words) and/or n-grams provides the most consistently reliable results in authorship attribution problems (Martindale und McKenzie, 1995; Diederich et al., 2003; Burrows, 2002; Hoover, 2003b,a; Uzuner und Katz, 2005; Zhao und Zobel, 2005; Grieve, 2007; Koppel et al., 2007; Yu, 2008).“ (Jockers und Witten, 2010)

fizierung von Texten anhand der Muttersprache des Autors (Abschnitt 3.6.2) und die Untersuchung von Textpaaren, deren Autoren Zwillinge sind (Abschnitt 3.6.4).

Die Untersuchungen überspannen einen weiten Bereich verschiedenartiger Testkorpora. Die Ergebnisse erlauben somit in ihrer Gesamtheit recht allgemeine Aussagen über die Eigenschaften und die Performanz des Verfahrens. Wo es möglich ist, wird das hier vorgestellte Verfahren mit veröffentlichten Performanzwerten bestehender Algorithmen verglichen.

3.2 Die stilometrische Forschungslandschaft

Es folgt ein Überblick über die veröffentlichte stilometrische Forschung. Erst vor diesem Hintergrund wird es möglich sein, meinen eigenen Ansatz vergleichend einzuführen.

Stilometrie als wissenschaftliche Disziplin ist bereits über hundert Jahre alt. Die übliche Zitatliste beginnt mit dem Ende des 19. Jahrhunderts. In der Frühzeit der Stilometrie wurden wegen der im Vergleich zu heute dramatischen Ressourcenknappheit Methoden verwendet, die mit besonders leicht erfassbaren Informationen arbeiteten, beispielsweise der Wortlänge (Mendenhall (1887)[5] oder auch Sherman (1888), letzterer zitiert nach Rudman (1998)). Als Mendenhall sein oft und oft nicht ganz korrekt[6] zitiertes Werk schrieb, war er als Physiker noch guter Hoffnung, in der Wortlängenverteilung eines Autors ähnlich eindeutige Signaturen zu finden wie man sie damals schon für die Emissionsspektren verschiedener Elemente kannte.[7] Er sollte sich täuschen, solche Spektren existieren nicht, und seine Vision ist nach wie vor in weiter Ferne.

Die anfänglichen Beschränkung der nutzbaren Ressourcen bestehen nun nicht mehr. Seit der Verfügbarkeit rechenstarker Computer sind den verwertbaren Datenmengen und den einsetzbaren Verfahren keine merklichen Grenzen mehr gesetzt.

Allgemein zum Thema Stilometrie und insbesondere zur automatischen Autorenbestimmung gibt es eine Reihe aktueller und etwas älterer Übersichtsartikel (Holmes, 1994, 1998; Juola, 2006a; Grieve, 2007; Stamatatos, 2009; Koppel et al., 2009). Diese Übersichtsartikel versuchen nach Möglichkeit, Struktur in die große Vielfalt der veröffentlichen Arbeiten zu bringen.Über die Situation bis Ende der neunziger Jahre geben die Werke von Holmes (1994, 1998) einen fundierten Überblick. Sein Artikel von 1994 ist eine lebendige Beschreibung vor allem der frühen geschichtlichen Entwicklung der Stilometrie. Der Beitrag von 1998 beschäftigt sich stärker mit den technischen Wirkungsweisen der dargestellten Methoden. Der damaligen Forschung entsprechend legt er das Gewicht

[5] Holmes (1998) verfolgt seine Idee sogar zurück bis 1851 (de Morgan, 1882).

[6] Weder zählte er Satzlängen wie beispielsweise von Diederich et al. (2003); Koppel et al. (2009) beschreiben. Und mit Shakespeare, wie von Stamatatos (2009) berichtet, beschäftigt er sich erst in Mendenhall (1901).

[7] „By the use of the spectroscope, a beam of [...] light is analyzed. [...] So certain and uniform are the results of this analysis, that the appearance of a particular spectrum is indisputable evidence of the presence of the element to which it belongs.

In a manner very similar, it is proposed to analyze a composition by forming what may be called a ‚word-spectrum' [...] If, now, it shall be found that with every author, as with every element, this spectrum persistst in its form and appearance, the value of the method will be at once conceded." (Mendenhall, 1887, S. 238)

auf *vocabulary-richness*-basierte Ansätze und ähnliche, relativ einfach strukturierte Verfahren. Multivariate Ansätze werden hier noch relativ kurz beschrieben, ihr kommender Siegeszug ist deutet sich aber bereits an. Das folgende Jahrzehnt wird von Stamatatos (2009) beschrieben. Das Hauptaugenmerk aller drei Übersichten liegt zwar auf der Autorenbestimmung, sie bieten aber auch einen Einblick in das allgemeinere Feld der Stilometrie. Die übrigen beiden Artikel (Grieve, 2007; Koppel et al., 2009) sind keine reinen Überblicksartikel, enthalten aber breite und fundierte Beschreibungen des Forschungsstandes.

Die inhaltliche Diversität auf dem Feld der Stilometrie ist wesentlich überschaubarer als in den Arbeiten zur *Morphologischen Induktion* (MI), die ich in 2.4 referiert habe. Bereits die Fragestellung der Stilometrie ist vergleichsweise einfach: *Wie und wie gut können Texte nach gewissen externen Variablen klassifiziert werden?* Demgegenüber war für die Darstellung der MI die Grundlegung eines theoretischen Unterbaus notwendig, um die Fragestellung überhaupt eindeutig formulieren zu können.

Ich teile die Forschungsansätze nach zwei Kriterien ein: Das erste Kriterium ist die Antwort auf die Frage, welche Daten herangezogen werden und in welche Strukturen sie überführt werden. Das zweite Kriterium ist die verwendete Klassifikationsmethode.

Die Datengrundlage der Stilometrie. *Stilometrie* ist Textklassifikation. Ursprüngliche Datenquelle wird daher immer ein Text sein. Die einfachste Repräsentation der Daten ist der Text als reine Zeichenkette. Mit einer einzigen langen Zeichenkette an sich lässt sich noch kein Text klassifzieren. Dazu muss sie erst in eine Menge an kleineren Zeichenketten zerlegt werden. Teahan (2000); Clement und Sharp (2003); Tsur und Rappoport (2007) sind drei Beispiele für Arbeiten auf dieser Datengrundlage. Trotz der strukturellen Einfachheit von *Zeichen-n*-Grammen ermöglichen sie effektive stilometrische Klassifikationen (Stamatatos, 2009, 542). In so gut wie allen relevanten Ansätzen wird die Information verwendet, welche Zeichenkette wie häufig im Text vorkommt.[8]

Bereits auf dieser Ebene aber zeigt sich ein grundlegendes Problem stilometrischer Forschung. Gewöhnlich werden nicht alle Zeichenketten in die Berechnung mit einbezogen, sondern der weitaus größte Teil von vornherein aussortiert. Üblicherweise geschieht das durch Begrenzung der maximalen Länge oder der minimalen Frequenz der Zeichenketten. Der Grund für eine solche Beschränkung bei der Auswahl der Daten ist zweierlei: Zum Einen ist es im allgemeinen wünschenswert, die verwendeten Modelle schlank zu halten und Speicher, Zeit und Rechenkraft zu sparen. Mit naiven Indexstrukturen kommt man an die Grenzen der Ressourcen, lange bevor alle Substrings eines Textes verarbeitet sind. Dieses technische Problem wird in der vorliegenden Arbeit durch den Einsatz von Suffixbäumen vermieden.

Der zweite Grund für die Einschränkung der verwendeten Daten auf kurze oder häufige

[8] Die einzige mir bekannte Ausnahme bildet der Ansatz von Benedetto et al. (2002a). Hier wird die praktische Komprimierbarkeit mit Hilfe des Standard-Algorithmus LZ77 zur Berechnung der Textähnlichkeit herangezogen. Einerseits werden so unbegrenzt lange Wiederholungen berücksichtigt, wie in der vorliegenden Arbeit. Andererseits spielen Frequenzen keine Rolle. Um die Berechtigung dieses Ansatzes wurde eine heftige Diskussion geführt (Goodman, 2002; Benedetto et al., 2002b; Khmelev und Teahan, 2003; Benedetto et al., 2003)

Zeichenketten liegt tiefer: Die Statistik der Zeichenketten eines Textes enthält nicht nur Informationen über Variablen, die den Stilometriker interessieren: Identität der Autorin oder des Autors, sein oder ihr Geschlecht, Alter oder ähnliches. Andere Informationen liegen sogar wesentlich offener zu Tage, allen voran über das Thema eines Textes: Ein Text, der die Wörter „Higgs-Teilchen“ und „Boson“, enthält, wird sich mit Teilchenphysik beschäftigen, ein Text mit den Wörtern „Ritter“ und „Prinzessin“, erzählt wohl ein Märchen. Dieses zweite Beispiel macht die Interaktion mit einer weiteren Variable offenbar, die ebenfalls in der Lage ist, die Oberflächenfrequenzen eines Textes stark zu beeinflussen: Dem *Genre*. Das Beispiel suggeriert, dass es *Genres* gibt, die eine spezielle Verteilung der Inhaltswörter nach sich ziehen können. Aber auch für die Funktionswörter kann ein solcher Effekt erwartet werden: Ein Roman oder eine Autobiographie wird mehr Personalpronomina enthalten als ein wissenschaftlicher Artikel. Das immanente Ziel der Stilometrie ist nun aber das dingfest machen des Einflusses der erwähnten stilometrischen Variablen, unabhängig von *Topic* (und *Genre*).

Es gibt die starke Tendenz, den Teil der Daten zu vermeiden, auf den das *Topic* eines Textes entscheidenden Einfluss hat. In Bezug auf Zeichenketten schließt das vor allem die langen und selteneren aus.[9] Problematisch kann es dabei sein, dass empirisch etabliert ist, dass *Genre*-Effekte im allgemeinen ebenfalls stärker sind als der persönliche Stil eines Autors.[10] Das *Genre* als Einflussgröße wird zu häufig nicht gesondert betrachtet. Ein Beispiel für die mangelnde Trennung des Einflusses von *Topic*, *Genre* und Autorschaft ist Granados et al. (2008). In dieser Arbeit werden englischsprachige Texte nach Autorenschaft klassifiziert ohne dass darauf eingegangen wird, dass sich neben den Autoren auch die Themen, das Genre, die Entstehungszeit und die Originalsprache erheblich unterscheiden. Die einführenden Bemerkungen oben lassen die Konzentration auf das *Topic* als Störvariable auch deshalb besorgniserregend scheinen, da das *Genre* einen schwer einschätzbaren Einfluss auf die Statistik der Funktionwörter haben könnte. Dies ist gerade der Teil der Daten, mit denen meist *Stilometrie* betrieben wird, um dem Einfluss des *Topic* aus dem Weg zu gehen. Die praktische Relevan derartiger Fragen wird in Abschnitt 3.6.4 untersucht.

Oberhalb der reinen Zeichenkette ist tokenisierter Text das nächst abstraktere Textniveau, das betrachtet werden kann. Entsprechend ergeben sich Statistiken der Oberflächenwortformen von Texten. Auch hier lassen sich Ketten betrachten, typischerweise bis zur maximalen Länge 3, da spätestens dann die sehr geringe Frequenz der einzelnen n-Gramme zu Problemen führt (vgl. Stamatatos, 2009, 541). Auch hier wird oft versucht, Wechselwirkungen der *stilometrischen Variablen* mit dem *Topic* des Textes zu umgehen, indem man den Teil der Daten ausschließt, in dem dessen Einfluss hauptsächlich ver-

[9] (Stamatatos, 2009, 545): „The most important criterium for selecting features in authorship attribution tasks is their frequency. In general, the more frequent a feature, the more stylistic variation it captures.“

[10] „Genre effects generally will supersede authorial features in the discrimination process.“ (Holmes, 1998); „analyses in this study have shown that time or genre effects are often so marked that they can partly mask authorship.“ (Forsyth et al., 1999); „texts in different registers or text types by one author may differ more than texts written by different authors in the same text type.“ (Baayen et al., 1996)

mutet wird. Nun steckt das *Topic* wie die Beispiele oben schon suggerieren hauptsächlich in den so genannten inhaltstragenden Wörtern (*content words*). Vermeidet man sie, so die Hoffnung, hat man sich unabhängig gemacht vom konkreten Inhalt des Textes. Übrig bleiben die so genannten *Funktionswörter* (*function words*), denen weniger eine lexikalische Bedeutung als eine grammatische Funktion zukommt.[11] Man nimmt an, mit ihnen eine Signatur zu besitzen, die vom *Topic* unabhängig ist.

Im allgemeinen sind die inhaltstragenden Wörter auch die selteneren, so dass eine Beschränkung auf die kurzen, häufigen Wörter demselben Zweck dient, den *Topic*-Effekt zu eliminieren. Drei Beispiele für ein solches Vorgehen aus verschiedenen Abschnitten der Stilometrie sind Mosteller und Wallace (1964); Burrows (1988); van Halteren (2008), die jeweils nur die häufigsten Typen auswerten.[12]

Diese Annahme einer Komplementarität[13] von *Topic* und *Stil* wird zum Beispiel bei Clement und Sharp (2003) explizit, wenn sie schreiben „[...] particles are generally seen [...] as not conveying meaning, but are considered to represent stylistic cues indicating authorship." Ähnlich eindeutig formuliert auch Stamatatos (2009): „The most common words [...] do not carry any semantic information. [...] they are topic-independent."

Während sich die stilometrische Forschungsgemeinschaft einig ist, dass *Funktionswörter* keine Korrelation mit dem *Topic* aufweisen, wird anders herum durchaus die Möglichkeit diskutiert, dass nicht nur die *Funktionswörter*, sondern auch die *inhaltstragenden Wörter* Hinweise auf *stilistische Variablen* enthalten:

> For example, one author may prefer to use the words *start* and *large*, where another may prefer *begin* and *big* [...]. Such patterns of lexical choice can be represented by modeling the relative frequencies of content words[...] (Koppel et al., 2009, 11)

Mit diesem Ansatz erzielen die Autoren tatsächlich gute Ergebnisse. Allerdings ist ihre Zielsetzung nicht eigentlich Stilometrie, sondern vor allem reine *Authorship* ***Attribution***. Konsequenterweise betrachten sie das *Topic* nicht unabhängig von *stilistischen* Variablen. Ihr Korpus ist so gelagert, dass eine mehr oder minder starke Korrelation von *Topic* und *Stil* wahrscheinlich ist.

Inwieweit die starke Annahme zutrifft, dass Informationen über *Stil* und *Topic* strikt auf den unterschiedlichen Ebenen der *Funktions-* und *Inhaltswörter* liegen, oder auch die schwächere, dass *Funktionswörter* nicht mit dem *Topic* wechselwirken, wurde meines Wissens bisher kaum systematisch untersucht. Neben der mangelnden Fundierung des üblichen Vorgehens existiert noch ein weiteres Argument, neue Wege zu suchen. So gibt es eine andere denkbare Strategie, mit dem Problem der starken Störvariablen *Topic* und *Genre* umzugehen, als ihren Einfluss durch eine Filterung der Daten zu unterdrücken: Die explizite und gemeinsame Untersuchung von *stilometrischen* Variablen und *Topic/Genre* mit dem eventuellen Ziel einer Entflechtung der verschiedenen Einflüsse. Eine der ganz wenigen Arbeiten, die die Wechselwirkungen verschiedener Variablen untersuchen, ist

[11] Beispiele sind die Elemente geschlossener Wortklassen wie *wie, und* und *wohl*.

[12] Im Detail betrachtet behält van Halteren (2008) die Wörter bei, die in mindestens 10% der Texte des untersuchten Korpus vorkommen.

[13] Stamatatos (2009, 544) verwendet sogar das Wort *orthogonal*.

Clement und Sharp (2003). Ich wende mich derartigen Fragen in Abschnitt 3.6.4 zu. Zu nennen ist in diesem Zusammenhang auch die Untersuchung von Golcher und Reznicek (2011). Dort wird gezeigt das enge Zusammenspiel von Textähnlichkeit, *Topic* und der Muttersprache von Lernern anhand des Lernerkorpus Falko (Lüdeling et al., 2008) untersucht.

Die nächst höhere Abstraktionsebene sind POS-annotierte Texte und Korpora (s. zB. Chaski, 2005; Koppel et al., 2005). Einerseits lassen sich n-Gramme von POS-Tags direkt zur stilometrischen Analyse heranziehen. Andererseits kann diese Information genutzt werden, die Funktionswörter von den inhaltstragenden Wörtern zu unterscheiden. Beide Möglichkeiten dienen wiederum dem Zweck, den Einfluss des *Topic* von vornherein zu unterdrücken.

Neben (Häufigkeiten von) Zeichenketten, (Häufigkeiten von) (Ketten von) (Funktions)wörtern und (Häufigkeiten von) (Ketten von) POS-tags gibt es relativ wenige weitere Arten von Informationen, die in bisherigen Arbeiten zu Rate gezogen werden. Zu nennen wäre vielleicht Koppel et al. (2005), die Fehler in Lernertexten mit einbeziehen. Baayen et al. (1996) sind die ersten, die syntaktische Bäume bzw. *rewrite rules* anstelle einer flachen POS-Annotierung verwenden. In Folge untersuchen auch Chaski (2001); Gamon (2004); Hirst und Feiguina (2007); Koppel et al. (2009) derartige Daten. Koppel et al. (2009) nennen noch weitere Beispiele für ein solches Vorgehen, sie beziehen sich aber auf POS-tags, Funktionswörter, Satzzeichen oder ähnlich flache Strukturen. Weiter auf die Details dieser Ansätze einzugehen wäre hier nicht von großem Nutzen, da syntaxbasierte Daten zu weit vom Fokus der vorliegenden Arbeit entfernt sind.

Alle beschriebenen Annotationsebenen, abgesehen vielleicht von der Tokenisierung, bringen sprachliches Wissen in die Analyse mit ein. Das macht die vorgeschlagenen Algorithmen vom Vorhandensein solchen Wissens abhängig, beschränkt ihre Anwendbarkeit also auf wenige Sprachen. Die in Kapitel 2 behandelte *Morphologische Induktion* ist eine Grundvoraussetzung so gut wie jeder weiteren Verarbeitung natürlicher Sprache. Dort ist daher die Sprachunabhängigkeit von Algorithmen allein aus Sicht möglicher Anwendungen ein vordringliches Ziel. In der Stilometrie dagegen tritt es in den Hintergrund und wird kaum thematisiert.

Im Allgemeinen erfolgen alle Annotierungsschritte automatisch. Damit ist eine unvermeidliche Restfehlerquote verbunden: „[...] the more detailed the text analysis required for extracting stylometric features, the less accurate (and the more noisy) the produced measures“ (Stamatatos, 2009, 543). Es ist denkbar, dass dieses Rauschen eine stilometrische Klassifizierung unmöglich macht. Es hat sich erwiesen, dass das nicht der Fall ist: Stilometrie mit automatisch annotierten Daten ist ohne weiteres möglich. Das zeigen alle derartigen hier zitierten Arbeiten.[14]

Ein paar Worte zur Verteilung der beschriebenen Klassen von Daten. In Koppel et al. (2009) findet sich eine tabellarische Übersicht über 71 stilometrische Arbeiten der letzten 120 Jahre. Bald die Hälfte davon (29) ziehen Funktionswörter zu Rate, die damit die

[14] Allerdings ist mir keine Arbeit bekannt, die überprüft, inwieweit die Fehler der verwendeten Algorithmen systematisch mit den untersuchten *stilometrischen* Variablen variieren und ob derartige Phänomene einen Einfluss auf den Erfolg der Klassifizierung haben.

verbreitetste Datenquelle darstellen.[15] Nur etwa halb so viele (16) verwenden *Zeichenketten*. Aber auch Wortformen im Allgemeinen, POS-tags und ihre jeweiligen n-Gramme finden sich auf den ersten Plätzen.

Die eingesetzten Klassifikationsverfahren. Damit sind die grundlegenden Datenquellen stilometrischer Verfahren vorgestellt. Sie werden jeweils zusammen mit vielen unterschiedlichen Klassifikationsverfahren eingesetzt. Drei Gattungen lassen sich beschreiben.

Die erste und zugleich älteste Methode weist jedem Text eine Kennzahl zu und versucht die stilometrische Zuordnung durch den Vergleich dieser Kennzahlen zu erreichen. Das Eingangs zitierte Werk von Mendenhall (1887) ist ein klassisches Beispiel für ein derartiges Vorgehen. Die von Mendenhall verwendete Kennzahl ist die durchschnittliche Wortlänge. Yule setzt hierzu in seiner Arbeit von 1938 die Satzlänge ein. Viele der definierten Kennzahlen versuchen ein Maß für den Reichtum des verwendeten Wortschatzes darzustellen (vgl. Yule, 1944, 1968).

Derartige Maße werden auch in einigen neueren Arbeiten verwendet (Abbasi und Chen, 2008; Tweedie und Baayen, 1998; Baayen et al., 1996; Holmes und Forsyth, 1995, s. zB.). Diese gehören aber nicht zur eben dargestellten einfachen Klassifikationsmethodik. In Holmes und Forsyth (1995); Baayen et al. (1996) werden mehrere Wortschatzreichtumsmaße gemeinsam betrachtet. So berechnen Holmes und Forsyth (1995) sechs verschiedene derartige Maße für die einzelnen Texte und fassen sie zu Vektoren zusammen, die dann wiederum den Input multivariater Methoden bilden. Für Abbasi und Chen (2008) sind Wortschatzreichtumsmaße nur eine unter vielen betrachteten quantifizierbaren Eigenschaften von Texten. Tweedie und Baayen (1998) wiederum ist nicht eigentlich eine stilometrische Arbeit, sondern beschäftigst sich kritisch mit den bekannten Wortschatzreichtumsmaßen.

Der zweite grundlegende Klassifikationsansatz geht von einem Text*paar* aus und weist beiden Texten zugleich eine Kennzahl zu. Sie lässt sich als *Ähnlichkeit* der Texte oder als ihr *stilistischer* Abstand interpretieren. Die Klassifikation erfolgt anhand dieses Vergleichsmaßes. Mein eigener Ansatz ist dieser Klasse zuzuordnen.

Zur Illustration gehe ich anhand eines stark vereinfachten Beispiels auf eine Unterart derartiger Verfahren genauer ein. Die Texte seien in Trainings- und Testdaten unterteilt. In einer Standardsituation der *Automatischen Autorenbestimmung* bestehen die Trainingsdaten aus den bekannten Texten eines Autors, die Testdaten aus den zu klassifizierenden Texten. Die Trainingsdaten werden in einem *language model* kodiert. Was damit gemeint ist, sei an einem kurzen Beispiel demonstriert. Angenommen, der Trainingstext besteht nur aus der Zeichenkette `abrakadabra`. Ein *language model* könnte nun aus dem Frequenzspektrum der einzelnen *Zeichen* bestehen: $h(\texttt{a}) = 5$, $h(\texttt{b}) = 2$ etc bestehen. Eine andere Möglichkeit wäre die Häufigkeitsverteilung von Bigrammen oder höheren n-Grammen. Auch bedingte Verteilungen sind vorstellbar: $h(\texttt{b}|\texttt{a}) = 2/5$ etc. Ein solches *language model* kann als eine Blackbox gesehen werden, die Text mit gewis-

[15] Juola (2006a) schreibt: „[...] the idea of mining function words for cues to authorship has become a dominant theme in modern research.“

sen statistischen Eigenschaften produziert. Dazu sind nur die gezählten Häufigkeiten als Wahrscheinlichkeiten zu interpretieren. Die Gefahren eines solchen Vorgehens waren bereits in Abschnitt 2.4 (S. 31) Thema.

Die Klassifizierung der Testtexte geschieht nun über einen Vergleich mit dem *language model*. Das kann nach dem bereits erwähnten *maximum likelihood*-Prinzip geschehen: Es wird das Modell gewählt, für das die Wahrscheinlichkeit, den Testtext zu produzieren am größten ist. Auch der uns bereits aus der MI bekannte Begriff der *Entropie* findet Verwendung. Dahinter steht der folgende Gedanke: Je besser wir ein Ereignis vorhersagen können, desto geringer ist seine Entropie (vergleiche die Ausführungen auf Seite 2.4ff). Sollen nun zwei Texte A und B verglichen werden, so wird wieder das aus A gewonnene *language model* verwendet, um B vorherzusagen. Je wahrscheinlicher Text B unter der aus A verwendeten Information ist, desto besser die Vorhersage und desto geringer ist die Entropie.[16] Nimmt man die Interpretation der berechneten Größen als wirklichen *Wahrscheinlichkeiten* und *Entropien* wiederum nicht allzu ernst, bleibt folgendes Kernverfahren übrig: Die grundlegenden Häufigkeitsdaten von Texten werden in eine summarische Repräsentation gebracht. Zwei Repräsentationen werden verglichen, indem aus beiden eine einzige Zahl berechnet wird. Zur Klassifizierung wird sie maximiert oder minimiert.

Da der Ansatz von Keselj und Cercone (2004) ein gut darstellbares Beispiel dieser Gruppe ist und weil es Vergleichspunkte zu meinem eigenen Algorithmus vergleichen lässt, referiere ich ihn hier kurz. Das dort verwendete Maß für die Unterschiedlichkeit zweier Texte T_1 und T_2 ist

$$d(T_1, T_2) = \sum_{s \in M} \left(\frac{2\left(f_{T_1}(s) - f_{T_2}(s)\right)}{f_{T_1}(s) + f_{T_2}(s)} \right)^2$$

Hier ist s ein *Zeichen-n*-Gramm und M enthält die N häufigsten n-Gramme beider Texte. Je größer d, desto größer der stilistische Abstand beider Texte. Ich komme im folgenden Abschnitt 3.3 auf dieses Maß zurück und vergleiche es explizit mit einer Variante meines eigenen Ansatzes.

Ebenfalls erwähnenswert ist in diesem Zusammenhang wieder die Arbeit von Teahan (2000). Er verwendet ein aus dem hoch effizienten Kompressionsalgorithmus PPM abgeleitetes *language model*. Sein Ähnlichkeitsmaß ist entropiebasiert. Siehe auch die Bemerkungen zu Teahan (2000) in den Abschnitten 3.3 und 3.5.

Für die bisher besprochenen Klassifikationsmethoden bestehen die zu Rate gezogenen Daten fast ausschließlich aus dem Text als Zeichenkette, allenfalls aus dem tokenisierten Text. Das ändert sich, wenn wir zur dritten grundlegenden Methode übergehen.

Hier ist der Ausgangspunkt eine multidimensionale Vektorrepräsentation der Dokumente. Die Dimensionen können sehr verschiedene Entitäten repräsentieren. Übliche Möglichkeiten sind Zeichenketten, Wortformen oder POS-tags, bzw. die jeweiligen n-Gramme. Diese Dimensionen werden als *features* bezeichnet. Die Werte in den einzelnen Dimensionen sind im Normalfall die Zahl der entsprechenden Vorkommen. Unser Train-

[16] Die enge Verwandtschaft der beiden Kriterien *maximum likelihood* und *minimum entropy* wird bereits in Kriz und Talacko (1968) diskutiert.

ingstext `abrakadabra` könnte so beispielsweise über einen Vektor der Art $(5, 2, 2, 1, 1)$ repräsentiert werden, dessen Dimensionen für die Vorkommen der Buchstaben `a`, `b`, `r`, `k`, `d` stehen. So verwandelt sich ein Dokument in einen Punkt in einem vieldimensionalen Raum.

Die letztendliche Klassifikation besteht darin, die im Raum verstreuten Punkte zu möglichst klar umrissenen Mengen zusammenzufassen.

Mosteller und Wallace (1964, 1984) sind die ersten Autoren, die in diese Gruppe eingegliedert werden können, auch wenn die räumliche Metapher hier noch fehlt. Sie legen den Grundstein sowohl für die Hinwendung der Stilometrie zu multivariaten Methoden, als auch zur Konzentration auf Funktionswörter. Sie sind auch die ersten, die Autorenbestimmung anhand der *Federalist Papers* betreiben. Dieses Korpus wird an anderer Stelle noch eine Rolle spielen und dort detaillierter beschrieben (Abschnitt 3.6.3). Hier soll genügen, dass es sich um 85 Aufsätze zweier Hauptautoren handelt, wobei die Autorenschaft von 12 Texten als nicht gesichert gilt. Ausgehend von den unterschiedlichen Frequenzen dieser Wörter bei den beiden Autoren verwenden sie das Bayes'sche Theorem, um die umstrittenen Aufsätze zuzuordnen. Diese Arbeit verdient auch Beachtung, weil sie als eine der einzigen wirklich fundierten Gebrauch von einem scharf umrissenen Wahrscheinlichkeitsbegriff macht.[17]

Ein weiteres relativ frühes Beispiel für ein derartiges Verfahren ist die Arbeit von Bosch und Smith (1998), die sich auch explizit an Mosteller und Wallace (1964, 1984) orientieren: Die Dimensionen (*features*), von denen Bosch und Smith (1998) ausgehen, sind die Frequenzen von 70 Funktionswörtern. In den von diesen Frequenzen aufgespannten 70-dimensionalen Raum ziehen sie Hyperebenen ein, um die beiden Autoren des Datensatzes voneinander zu trennen und die umstrittenen Aufsätze zu klassifzieren. Sie wählen durch Minimieren der Fehlerquote eine möglichst kleine Menge an Wörtern, die in der Lage sind, zwischen den Autoren zu unterscheiden. Es ergeben sich die drei Wörter *are*, *our* und *upon*. Ihre Methode ist im Nachhinein betrachtet noch nicht ganz ausgereift, zeigt aber bereits wesentliche Grundzüge dieser Familie von Ansätzen.

Eine weitere Abstraktionsstufe wird durch die *Hauptkomponentenanalyse*, oder PCA (*Principal Component Analysis*) erreicht (Burrows, 1987, 1988, 1992; Holmes und Forsyth, 1995; Baayen et al., 1996; Forsyth et al., 1999; Holmes et al., 2001; Gamon, 2004). Dieses Verfahren ist im Kern eine Koordinatentransformation. Die Koordinaten des ursprünglichen vieldimensionalen Raumes werden so gedreht, dass möglichst wenige der gedrehten Koordinaten (Hauptkomponenten) einen möglichst großen Anteil der Varianz in den Daten beschreiben. Die übrigen Koordinaten werden ignoriert. Beschränkt man sich auf die zwei wichtigsten Hauptkompomenten, kann man die Datenpunkte nun in einer Ebene graphisch darstellen. Eine objektive Klassifikation über diese visuelle Darstellung hinaus unterbleibt allerdings meist.

Diese Lücke wird mit der Einführung von Maschinenlernverfahren in die Stilometrie geschlossen. Maschinenlernverfahren gehen von den im Raum verteilten Dokumentvektoren bekannter Klassifikation aus, leiten aus diesen ein festes Klassifikationsverfahren

[17] Holmes und Forsyth (1995) geben eine zugängliche Darstellung der Monographien von Mosteller und Wallace (1964, 1984)

ab, das dann auf die Testdaten angewendet werden kann. Im Laufe der Zeit wurden recht unterschiedliche Lernverfahren eingesetzt. Einige Beispiele umfassen neuronale Netze (Tweedie et al., 1996), Support Vector Machines (SVM) (Diederich et al., 2003; Fung, 2003; Baroni und Bernardini, 2006; Hirst und Feiguina, 2007; Koppel et al., 2009) oder *Bayesian regression* (Koppel et al., 2009), um nur

drei zu nennen.[18] Die bekannteste dieser Methoden ist SVM,[19] der derzeitige *de facto* Standard (SVM wurde beispielsweise vom Gewinner des von Juola et al. (2006) veranstalteten *Authorship Attribution Contests* eingesetzt). Diese Methode verfolgt einen Ansatz, der im Rahmen von Bosch und Smith (1998) beschriebenen Linie folgt, diesen Weg aber weitergeht: Die eingezogene Hyperebene wird algorithmisch so optimiert, dass sie die bekannten Positivbeispiele möglichst klar separiert.

In der Forschungsgemeinschaft herrscht weitgehend Einigkeit, dass die modernen maschinellen Lernverfahren gegenüber den traditionellen Verfahren im Vorteil sind. Die Überlegenheit der maschinellen Lernverfahren könnte daher rühren, dass hier im allgemeinen Informationen aus allen Texten miteinander verknüpft werden, nicht nur von zwei jeweils miteinander verglichenen. Das hier vorgestellten Verfahren hat ähnliche Eigenschaften (Abschnitt 3.4).

In der vorangegangenen Besprechung der verwendeten Methoden fehlen quantitative Angaben zur Performanz der verschiedenen Ansätze fast vollkommen. Die Ursache liegt in der allgemeinen Struktur des Forschungsfeldes. Leider hat sich nichts grundsätzliches geändert, seit Juola (2006b) schrieb:

> The current state of the art is an ad hoc mess of disparate methods with little cross-comparison to determine which methods work and which don't. Or more accurately, because they all work at least reasonably well ([...]), which methods work the best.

Diese unbefriedigende Situation ist bereits seit längerem bekannt. Schon Ende der 90er Jahre beschrieb Rudman in einem viel beachteten Artikel den Zustand des Forschungsbereichs mit harschen Worten:

> studies governed by expediency; a lack of competent research; flawed statistical techniques; corrupted primary data; lack of expertise in allied fields; a dilettantish approach; inadequate treatment of errors. (Rudman, 1998)

Es ist fraglich, ob man nach wie vor so weitgehende Kritik üben muss, aber das Problem, das Juola im ersten Zitat anspricht, besteht fort und hat klare Gründe: Die Performanz eines bestimmten stilometrischen Ansatzes wird mit Sicherheit abhängen von der Textlänge, vom Genre, von der Art der stilometrischen Fragestellung, von der Topic-Homogenität des Datensatzes und vielleicht vielen anderen Faktoren. Zwei Ansätze sind daher nur dann vergleichbar, wenn sie anhand derselben Fragestellung und anhand derselben Daten getestet wurden.

[18] Für weitere s. zB. Jockers und Witten (2010).
[19] Für einen kurzen übersichtlichen und linguistisch relevanten Einstiegstext siehe Baayen (2008, S.160ff).

Daraus leitet sich die Notwendigkeit eines allgemeinen Benchmarkkorpus ab, anhand dessen die Community die Qualität von Neuveröffentlichungen messen kann. Dennoch beruht die überwiegende Zahl der Veröffentlichungen auf *ad hoc*-Korpora (Koppel et al., 2009). Dieses Defizit ist schon vor längerer Zeit erkannt worden, unter anderem von Forsyth (1997). Dort werden Designprinzipien für ein Benchmarkkorpus vorgeschlagen und ein fertiges Korpus wird vorgestellt. Mir ist keine Arbeit bewusst, in der es verwendet worden wäre. Auch Juola (2004) hat ein kleineres, als Benchmark nutzbares Korpus vorgestellt, das meines Wissens aber ebenfalls keine weitere Verbreitung gefunden hat. Ich werde es in Abschnitt 3.5 verwenden, um Varianten meines Algorithmus zu vergleichen. Eine vollständigere Liste der im Laufe der Zeit vorgeschlagenen Evaluationskorpora findet sich bei Stamatatos (2009, 552). Alleine die Länge dieser Liste unterstreicht das grundlegende Problem, dass sich die Forschungsgemeinde bisher auf keinen Standard einigen konnte.

Bemerkenswerterweise treffen wir hier also auf genau dieselbe beklagenswerte Situation wie bereits bei der Besprechung der Literatur zur *Morphologischen Induktion* (s. Seite 28). Trotz jahrzehntelanger Forschung gibt es wenig Möglichkeiten, die Fülle der vorgeschlagenen Ansätze umfassend und objektiv aneinander zu messen.

In letzter Zeit allerdings gibt es viel versprechende Ansätze, die verschiedenen Methoden objektiv und systematisch zu vergleichen. Grieve (2007) stellt die Frage, welche der vielen im Laufe der Zeit zu Rate gezogenen Textähnlichkeitsmaße[20] am besten in der Lage sind, zwischen verschiedenen Autoren zu unterscheiden. Jockers und Witten (2010) untersuchen die komplementäre Frage, welche der modernen Klassifikationsmethoden am effektivsten sind.[21] Koppel et al. (2009) wiederum wenden sich dem Vergleich von Kombinationen aus Datengrundlage und Klassifizierungsverfahren[22] zu.

Aber auch hier ist es so, dass alle drei Arbeiten unterschiedliche Fragen bearbeiten und Vergleiche auf unterschiedlichen Ebenen durchführen. Auch verwenden alle ihre eigenen Korpora, insgesamt 5, von denen nur eines – die *Federalist Papers* – auch sonst häufiger verwendet wird (Zählt man die in Koppel et al. (2009) gegebene Tabelle aus, ist es sogar das am am häufigsten verwendete Korpus). Genau dieses Korpus allerdings ist als Benchmarkkorpus nicht besonders geeignet, weil es äußerst homogen ist und die zugrundeliegende Fragestellung eigentlich zu leicht ist.[23] Da es aufgrund seiner weiten Verbreitung eine gewisse Vergleichbarkeit gewährleistet, untersuche allerdings ich es hier auch (Abschnitt 3.6.3). Abgesehen von seiner Popularität wird es auch dazu dienen, eine interessante Variante des Algorithmus einzuführen.

Zusammengefasst: Auch fortschrittliche Projekte, die es unternehmen, Vergleiche zu ziehen, führen zu Ergebnissen, die wiederum untereinander schwer zu vergleichen sind. Die Möglichkeit, echte Objektivität zu schaffen, indem sich die Community auf ein

[20] Er vergleicht 39 Maße.

[21] Sie untersuchen 5 verschiedene maschinelle Lern- und Clusteringverfahren anhand der *federalist papers*.

[22] Hier wurden jeweils 5×5 Möglichkeiten getestet.

[23] Es gibt hierzu auch die gegenteilige Meinung. Vgl. z.B. Holmes und Forsyth (1995, 114), die aufgrund der großen Ähnlichkeit des Stils von Madison und Hamilton argumentieren, dass die *federalist papers* eine „ernstzunehmende Wahl für ein Testproblem“ darstellen. Die geringe Varianz der Ergebnisse, die sich in der Vielfalt der Arbeiten auf diesem Korpus zeigt, hat diese Aussage aber empirisch widerlegt.

verbindliches Evaluationskorpus einigt, wird nach wie vor verschenkt. Für spezielle Fragestellungen wird man immer spezielle Korpora benötigen, aber nichts spricht gegen eine entsprechend erweiterbare allgemeine Testsuite.[24]

Ein weiterer Kritikpunkt an großen Teilen der einschlägigen Forschungsliteratur ist das fast völlige Fehlen von Konfidenzintervallen für die eigenen Ergebnisse und die Durchführung von Signifikanztests beim Vergleich der eigenen Zahlen mit konkurrierenden Ansätzen. Zu wenige Autorinnen und Autoren sichern sich dagegen ab, dass die beobachteten Performanzunterschiede wenigstens auf dem untersuchten Korpus dem Zufall entstammen. Ein willkürlich ausgewähltes Beispiel stellt die Arbeit von Jockers und Witten (2010) dar. Die Autoren verwenden 70 Testtexte aus den *federalist papers* und zählen die Klassifikationsfehler des Algorithmus. Aufgrund dieser Fehlerzählungen bewerten sie die getesteten Methoden als unterschiedlich gut. Nimmt man aber die einzige Methode heraus, die deutlich schlechter als alle anderen, bleiben Fehlerzahlen zwischen 0 und 4 übrig. Der Fisher-Test verneint einen signifikanten Unterschied ($p = 0.12$). Dass die Verwendung des Fisher-Tests hier wegen des immer gleichen Testkorpus angreifbar ist, ist kein Argument gegen die Berechtigung der Frage: Wäre es nicht umso notwendiger, die Signifikanz und Stabilität der gemessenen Unterschiede sorgfältig zu diskutieren? Viele andere Arbeiten verwenden *cross validation*, ohne aber weiter auf die auftretenden Varianzen einzugehen.

3.3 Eine Familie von Textähnlichkeitsmaßen

Nun möchte ich meinen eigenen Ansatz vorstellen um ihn anschließend – nach und nach – in Bezug auf die gerade referierten existierenden Ansätze einzuordnen. Das grundlegende Verfahren ist schnell beschrieben. Ausgehend von den vollständigen Substringhäufigkeiten der beiden zu vergleichenden Texte wird eine Zahl berechnet, die sich als ihre Ähnlichkeit interpretieren lässt. Damit ordnet sich die Methode eindeutig in die Reihe der Ansätze ein, die kein maschinelles Lernverfahren verwenden, sondern die Dokumente aufgrund eines einfachen Zahlenvergleiches kategorisieren.[25] Ausgehend von einer einfachen Grundform werde ich eine Reihe von Varianten dieses Ähnlichkeitsmaßes vorstellen und motivieren. Im folgenden Abschnitt (3.4) wird ein wichtiger Normierungsschritt eingeführt, bevor dann in 3.5 die vorgestellten Maße empirisch verglichen werden.

Im Gegensatz zu vielen anderen Ansätzen (vgl. den vorhergehenden Abschnitt) schließt der vorliegende Ansatz nicht von vornherein die langen, seltenen Zeichenketten aus, sondern bezieht sämtliche Daten mit ein.

Wie bereits auf Seite 133 erwähnt gibt es meiner Kenntnis nach keinen weiteren Forschungsansatz, der die volle Menge der hier verwendeten Daten für stilometrische

[24] Über die Gründe für diesen lang anhaltenden strukturellen Mangel der aktuellen stilometrischen Forschung kann man nur spekulieren. Es ist aber abzusehen, dass es mit solch einem Korpus schnell schwierig wäre, Methoden zu präsentieren, die eine Verbesserung im Vergleich zu allen anderen Arbeiten darstellen, dh. Rekorde brechen könnten. Entsprechend schwierig könnte es dann sein, Veröffentlichungen dieser Art zu präsentieren. Genau dies ist aber der Tenor vieler Arbeiten.

[25] Obwohl in Abschnitt 3.6.3 doch ein maschinelles Lernverfahren (SVM) eingesetzt wird, aber dies mehr am Rande.

Untersuchungen verwendet. Selbst in Zhang und Lee (2006) dienen Suffixbäume, die diese Information ja enthalten, lediglich der Filterung der anscheinend am effektivsten nutzbaren Zeichenketten.

Dieser Vernachlässigung stehen empirische Erkenntnisse gegenüber, dass es sich lohnt, bei den längeren und selteneren Zeichenketten genauer hinzusehen. In der Einleitung (Kapitel 1) werden Arbeiten zitiert und referiert, die empirisch nachweisen, dass die Korrelationen zwischen zwei Textstellen mit ihrem Abstand so langsam schwächer werden, dass kein typischer Abstand mehr angenommen werden kann, oberhalb dessen Korrelationen keine Rolle mehr spielen (Schenkel et al., 1993; Amit et al., 1994; Ebeling und Pöschel, 1994; Ebeling und Neiman, 1995; Ebeling et al., 1995; Montemurro und Pury, 2002). Verwendet man von vornherein nur n-Gramme einer bestimmten Länge, so könnte dies dazu führen, dass Information, die in den längeren Zeichenketten steckt, übersehen wird.

Es gilt nun, aus den Daten, die uns zur Verfügung stehen – alle Substringhäufigkeiten der zu untersuchenden Texte –, ein Maß für die Ähnlichkeit zweier Texte zu gewinnen. Es ist naheliegend, genau die Zeichenketten zu betrachten, die in beiden Texten auftreten, ihre Frequenzen miteinander zu kombinieren und die Beiträge der einzelnen Zeichenketten in geeigneter Weise aufzusummieren. Wie weiter unten genauer ausgeführt wird, bieten sich mehrere Varianten an, diese Summierung durchzuführen.

Seien T_1 und T_2 zwei Texte. Das zur Verfügung stehende Datenmaterial besteht aus den Frequenzen aller Substrings beider Texte. Die Zahl der Vorkommen eines Strings s in T_1 sei mit $F_{T_1}(s)$ bezeichnet, analog für T_2. Das zu definierende Maß S für die Ähnlichkeit beider Texte sollte drei heuristischen Forderungen genügen: Erstens gehen nur Substrings ein, die in beiden Texten vorkommen. Zweitens soll S mit den Frequenzen in beiden Texten streng monoton steigen. Drittens scheint es sinnvoll, diejenigen Zeichenketten besonders hoch zu bewerten, die in beiden Texten gleichzeitig häufig vorkommen. Eine strukturell einfache Definition, die diesen Anforderungen genügt, ist die Summe über alle Produkte $F_{T_1}(s)F_{T_2}(s)$ für alle Substrings s:

Definition 33 *Der lineare Ähnlichkeitsindex S_l für die Texte T_1 und T_2 ist definiert als*

$$S_l(T_1, T_2) = \sum_{\text{alle } s} F_{T_1}(s)F_{T_2}(s)$$

s läuft über die Menge aller möglichen Zeichenketten.

Im Rahmen des klassischen Vektorraummodells entspräche der *lineare Ähnlichkeitsindex* dem Skalarprodukt.

Die Summe in dieser Definition läuft über alle denkbaren Substrings, weil dies die mathematisch einfachste Formulierung ist. Allerdings leisten Substrings, die in einem der beiden Texte nicht vorkommen, keinen Beitrag zur Summe ($F_i(s) = 0$ für $i = 1$ oder 2). In der technischen Implementierung beschränkt sich das Programm natürlich von vornherein auf diejenigen Substrings, die in beiden Texten vorkommen. Andere erscheinen nicht im Suffixbaum (Abschnitt 1.2).

S_l hat strukturelle Ähnlichkeiten mit anderen in der Literatur vorgeschlagenen Maßen.

In Abschnitt 3.2 (Seite 140) wurde bereits der Ansatz von Keselj und Cercone (2004) vorgestellt, der auf dem Abstandsmaß

$$d(T_1, T_2) = \sum_{s \in M} \left(\frac{2\left(f_{T_1}(s) - f_{T_2}(s)\right)}{f_{T_1}(s) + f_{T_2}(s)} \right)^2$$

basiert. Die mathematischen Unterschiede sind offenkundig, aber man kann argumentieren, dass sie sich als nicht besonders bedeutend erweisen sollten: Wo in S ein Produkt zweier Frequenzen steht, wird in d die Differenz dieser Frequenzen eingesetzt. Dh. große Werte von S werden mit kleinen Werten von d korrespondieren. Dies stimmt überein mit der Interpretation von S als Ähnlichkeit und d als Abstand zweier Texte. Der entscheidende Unterschied ist die Menge der ausgewerteten Substrings: In S wird über alle Substrings summierte, M dagegen läuft explizit über die Menge der N häufigsten n-Gramme mit $3 \leq N \leq 5000$. Dies lässt zum einen einen großen Teil der Daten unanalyiert und führt zum anderen einen neuen Parameter (N) in den Algorithmus ein. Aber es gibt noch ein weiteres strukturelles Argument gegen die Repräsentation der Frequenzen selbst. Dieses richtet sich allerdings gegen beide Maße, das d von Keselj und Cercone (2004), und das hier vorgeschlagene S_l.

Die Häufigkeiten von Substrings in Texten sind von sehr unterschiedlicher Größenordnung, unter anderem in Abhängigkeit von ihrer Länge. Bereits in einem relativ kurzen Werk wie Bebel (2004a) überspannen sie 5 Größenordnungen.[26] Diese wohlbekannte Tatsache spielte auch schon im Rahmen der *Morphologischen Induktion* eine große Rolle und führte zum dort so bezeichneten *Abfallproblem*.

Daher ist zu erwarten, dass Definition 33 für effektive Stilometrie nicht die ideale Wahl ist. Die extrem hohen Frequenzen vor allem der kurzen Zeichenketten werden jeden Einfluss der für die reale Textähnlichkeit vielleicht mitentscheidenden selteneren Zeichenketten überdecken. Da die jeweils verwendeten Frequenzen bei den meisten verwendeten stilometrischen Verfahren linear eingehen, kann es sehr gut sein, dass sie bisher systematisch unterschätzt wurden. Was aber geschieht, wenn man dazu übergeht die Daten so zu transformieren, dass Frequenzen verschiedener Größenordnungen vergleichbar werden?

Diese Überlegung lässt die Verwendung des Logarithmus $\log(x)$ natürlich erscheinen, der Unterschiede für kleine x stärker sichtbar macht, während die Logarithmen für große x näher zusammenrücken.

Ein weiterer Hinweis, dass die Verwendung des Logarithmus günstig sein könnte, lässt sich aus dem ersten Teil der Arbeit ableiten, wo die Kombination der Logarithmen der *Vorhersagbarkeitsabfälle* konsistent die besten Ergebnisse hervorbrachte. Auch die Ergebnisse von Diederich et al. (2003) deuten darauf hin, dass logarithmisch transformierte Frequenzen für stilometrische Zwecke überlegen sind[27]. Diederich et al. (2003) verwenden *Support Vector Maschines* zur Klassifikation auf Grundlage von Worthäufigkeitsverteilungen.

Würde der Logarithmus nun aber auf das reine Frequenzprodukt $F_{T_1}(s)F_{T_2}(s)$

[26] Das `e` erscheint 63553 Mal, die meisten Zeichenketten dagegen kommen nur einmal vor.
[27] „Simple relative frequencies do much worse than the logarithmic version."

angewendet, so blieben Substrings, die in beiden Texten nur ein einziges Mal erscheinen ($F_{T_1}(s) = 1$ und $F_{T_2} = 1$), unberücksichtigt, da gilt: $\log(1 \cdot 1) = \log(1) = 0$. Darüber hinaus müsste man Strings, die in einem der Texte nicht vorkommen, explizit aus der Definition herausnehmen, da der Logarithmus von 0 nicht definiert ist. Beide Probleme können umgangen werden, wenn zum Produkt der Häufigkeiten eine Konstante hinzugezählt wird. Setzen wir diese Konstante auf eins, ergibt sich:

Definition 34 *Der logarithmische Ähnlichkeitsindex S_{log} für die Texte T_1 und T_2 ist definiert als*

$$S_{log}(T_1, T_2) = \sum_{\text{alles}} \log\left(F_{T_1}(s)F_{T_2}(s) + 1\right)$$

Der Beitrag hoher Frequenzen bleibt durch das Addieren von 1 so gut wie unverändert. Die ganz kleinen werden ein wenig stärker gewichtet oder überhaupt erst berücksichtigt. Zeichenketten, die in einem der Texte nicht vorkommen, müssen wiederum nicht explizit ausgeschlossen werden, da $\log(0 \cdot a + 1) = 0$.

Da für alle Zahlen $a, b > 0$ gilt, dass $\log(ab) = \log(a) + \log(b)$, zerfallen die Summanden von S_{log} für den Grenzfall großer Frequenzen selbst in Summen, da dann $F_{T_1}(s)F_{T_2}(s) + 1 \approx F_{T_1}(s)F_{T_2}(s)$. Dies macht die Vorkommen eines Strings in den beiden verglichenen Texten unabhängig voneinander. Möchte man die Eigenschaft wieder herstellen, dass Zeichenketten, die in beiden Texten besonders häufig vorkommen, auch ein besonders hohes Gewicht bekommen, bietet sich folgende Modifikation an:

Definition 35 *Der multiplikativ logarithmische Ähnlichkeitsindex für die Texte T_1 und T_2 ist definiert als*

$$S_{mlog}(T_1, T_2) = \sum_{\text{alles}} \log\left(F_{T_1}(s) + 1\right) \log\left(F_{T_2}(s) + 1\right)$$

Die Verwendung des doppelten Logarithmus ist meines Wissens in der Stilometrie bisher einzigartig. Der Logarithmus an sich taucht aber durchaus auf und zwar in einer ganz bestimmten Gruppe von Arbeiten.

Ich habe mich bereits im vorigen Kapitel (2.4 29 ff.) eingehend mit dem Begriff der *Entropie*, seiner Geschichte und seiner Wirkung und Stellung in der Linguistik auseinandergesetzt. Auch in der Stilometrie taucht die *Entropie* immer mal wieder als Konzept auf. Die Motivation für diese Übernahme wird auch im Forschungsüberlbick zur Stilometrie (Abschnitt 3.2) diskutiert. Dort argumentiere ich, dass *Entropie* in der Stilometrie letztendlich als ein spezielles Textähnlichkeitsmaß interpretiert werden kann: Je ähnlicher die Texte, desto geringer die Entropie.

Davon unabhängig ist es nicht wirklich ein Zufall, dass man in Stilometrie und Morphologischer Induktion immer mal wieder auf identische oder eng verwandte Konzepte und Ideen trifft. Auf Teahan (2000), der in einer Arbeit beide Themenkomplexe mit demselben Verfahren untersucht wurde bereits in der Einleitung (Kapitel 1) hingewiesen. Abera auch allgemein analysieren vor allem die Verfahren, die sich an der Oberfläche des Textes orientieren, notwendigerweise strukturell sehr ähnliche Daten. Damit liegt eine gewisse Ähnlichkeit der Strategien nahe.

Bereits in der Diskussion der *Entropie* im vorigen Kapitel wird beschrieben, dass man zu ihrer Berechnung Wahrscheinlichkeiten braucht. Diese müssen auch hier geschätzt werden, auch hier aus Häufigkeiten. Dass dies ein gefährliches Vorgehen ist, wird dort ebenfalls schon gezeigt. Das Argument basiert im Wesentlich auf begrifflichen Problemen bei dieser Uminterpretation von Häufigkeiten in Wahrscheinlichkeiten. Nun gehe ich kurz auf ein Beispiel ein, das die daraus erwachsenden praktischen Probleme illustriert:

Clement und Sharp (2003) betreiben Autorenbestimmung mit n-Grammmodellen. Die n-Grammmodelle werden aus den Trainingstexten der verschiedenen Autoren gewonnen. Die eingehenden n-Grammhäufigkeiten werden als Wahrscheinlichkeiten interpretiert. Entsprechend wird ein Text demjenigen Autor zugeschlagen, dessen n-Grammmodell ihn mit der *größten Wahrscheinlichkeit* produziert haben könnte. Dies ist das bereits mehrfach erwähnte *Maximum Likelihood*-Verfahren.

Über die Länge des Textes werden die n-Gramme als unabhängige Ereignisse begriffen. Die Wahrscheinlichkeit unabhängiger Ereignisse berechnet sich aus dem Produkt der Einzelwahrscheinlichkeiten. Dieses Produkt, das zwangsläufig bei der Verknüpfung unabhängiger Ereignisse auftritt, ist eine direkte und zwingende Folge daraus, dass die Häufigkeiten als Wahrscheinlichkeiten interpretiert werden.

In diesem Produkt erscheinen nun auch Substrings, die zwar im Testtext, nicht aber im Trainingstext auftreten. Damit ist ihre auch relative Häufigkeit 0. Dasselbe gilt auch für die Auftretenswahrscheinlichkeit dieser Zeichenketten, da diese über die relative Häufigkeit im Trainingstext abgeschätzt wird. Damit diese zu Null geschätzten Einzelwahrscheinlichkeiten die Gesamtwahrscheinlichkeit nicht ebenfalls Null werden lassen, sind Clement und Sharp gezwungen, *smoothing*-Parameter einzuführen. Diese verkomplizieren das Modell und bringen Willkürlichkeiten mit sich.

In diesem Sinne zwingt ein ungerechtfertigtes theoretisches Fundament leicht in ein zu enges Korsett ohne den Erkenntnisgewinn einer echten Theorie.

Auch die Einführung des Entropiebegriffs in die Stilometrie gibt einen solchen festen Rahmen vor, da man sich damit auf eine sehr feste mathematische Form festgelegt hat ($\sum p \log p$), ohne, dass dahinter eine etablierte Theorie stehen würde, die eine solche Festlegung rechtfertigt.

Ich vermeide es daher konsequenterweise, Häufigkeiten als Wahrscheinlichkeiten zu interpretieren. Ebenso unterlasse ich es, meinen Definitionen ein theoretisches Fundament zu geben, da dieses beim derzeitigen Stand der Forschung leicht zu einem bloßen Label wird („Entropie“).

Dennoch werde ich in die Untersuchung Varianten von S einbeziehen, die in ihrer mathematischen Form den in der Literatur verwendeten Entropiefunktionen weitestgehend ähneln, um vergleichende Schlüsse ziehen zu können (Abschnitt 3.5, Seite 160).

Eine einflussreiche Arbeit (Teahan, 2000) wendet *cross entropy* auf verschiedene stilometrische Fragestellungen an: Im Allgemeinen ist *cross entropy* definiert als

$$\mathrm{H}(p, q) = -\sum_x p(x)\ \log q(x)$$

p und q bezeichnen hier zwei Wahrscheinlichkeitsverteilungen. Auch Juola und Baayen

(2005) verwenden dieses Maß.

Um Maße zur Verfügung zu haben, die sich möglichst eng an diesen öfters verwendeten Begriff der *cross entropy* anlehnen, definiere ich zwei weitere Maße:

Definition 36 *Der links logarithmische Ähnlichkeitsindex für die Texte T_1 und T_2 ist definiert als*

$$S_{llog}(T_1, T_2) = \sum_{\text{alles}} F_{T_1}(s) \left(\log \left(F_{T_2}(s) + 1\right)\right)$$

Analog ist der rechts logarithmische Ähnlichkeitsindex definiert als

$$S_{rlog}(T_1, T_2) = \sum_{\text{alles}} F_{T_2}(s) \left(\log \left(F_{T_1}(s) + 1\right)\right)$$

Singh und Gorla (2007)[28] machen gegenüber der klassischen *cross entropy* den Einwand, dass diese bei Vertauschung der beiden Verteilungen nicht symmetrisch ist. Daher definieren sie eine Entropievariante, die sie *symmetric cross entropy* nennen:[29]

$$sim(T_1, T_2) = \sum_{\text{alles}} p_1(s) \log p_2(s) + p_2(s) \log p_1(s) \tag{3.1}$$

p_1 und p_2 werden von Singh und Gorla (2007) als *distributions* bezeichnet, was wohl mit der *Wahrscheinlichkeit* des Vorkommens von s in den beiden Texten T_1 und T_2 bedeuten soll. Eine genaue Definition geben die Autoren nicht.

Folgende Definition versucht ein strukturell ähnliches Maß auf Grundlage unserer Daten nachzubauen:

Definition 37 *Der symmetrisch halblogarithmische Ähnlichkeitsindex S_{shlog} für die Texte T_1 und T_2 ist definiert als*

$$S_{shlog}(T_1, T_2) = \sum_{\text{alles}} F_{T_1}(s) \left[\log \left(F_{T_2}(s) + 1\right)\right] + F_{T_2}(s) \left[\log \left(F_{T_1}(s) + 1\right)\right]$$

Der einzige Unterschied ist, dass in Definition 37 absolute Häufigkeiten stehen, während in Gleichung 3.1 wohl relative Häufigkeiten gemeint sind.

Die vorgestellten Varianten von S erlauben es nun, zwei eng zusammenhängende Fragen empirisch zu klären: Welche Art von Daten erlauben eine sensiblere Messung der stilistischen Eigenschaften von Texten: die rohen Frequenzdaten oder ihre Logarithmen? Diese Frage klingt recht technisch, dh. vom linguistischen Standpunkt aus nicht sehr interessant. Man kann der Frage aber eine spannendere Form geben: Trifft die häufig gemachte Annahme (sh. Forschungsüberblick 3.2) zu, dass der Stil eines Textes sich maßgeblich an den häufigen Elementen auf Wortebene festmachen lässt? Erweist sich

[28] Diese Arbeit ist strengenommen nicht der Stilometrie zu zuordnen, es geht um die Differenzierung verschiedener Sprachen in mehrsprachigen Dokumenten. Die Probleme und Verfahren sind hier jedoch wiederum ähnliche.

[29] Die Notation ist ein wenig angepasst.

nämlich die logarithmische Darstellung der Daten als überlegen, wäre dies ein starker Hinweis auf die Bedeutsamkeit der längeren Ketten bzw. selteneren Phänomene.

3.4 Die Normierung von S

In den Definitionen 33 bis 37 stecken jeweils offensichtliche Textlängenabhängigkeiten: Je länger der Text, desto höher werden im Allgemeinen die beobachteten Frequenzen sein. Damit wachsen auch die verschiedenen S-Maße monoton an.

Auf den ersten Blick scheint folgende Strategie naheliegend. Sei die Länge eines Textes mit $L(T)$ bezeichnet. Angenommen, es gelingt, die Längenabhängigkeit von T_1 explizit in einer Funktion f zu beschreiben. Dann könnte man formulieren:

$$S(T_1, T_2) = f(L(T_1))S'(T_1, T_2)$$

Hinter dieser Formulierung steckt die Idee, dass S' nun nicht mehr von der Länge von T_1 abhängt. Würde T_1 homogen genug verlängert, zum Beispiel durch Verdoppelung, ändert sich nur $f(L(T_1)$. Vor allem für die Fälle, in denen man $S(T_1, T_2)$ als symmetrisch gegenüber einer Vertauschung von T_1 und T_2 annehmen kann, beschreibt f auch die Abhängigkeit von S von $L(T_2)$. Man kann also weiter schreiben:

$$S(T_1, T_2) = f(L(T_1))f(L(T_2))S_{norm}(T_1, T_2)$$

Der verbleibende Teil $S_{norm}(T_1, T_2)$ ist nun von der Länge beider Texte unabhängig. Man kann nun explizit nach dieser „normierten“ Version von S auflösen:

$$S_{norm}(T_1, T_2) = \frac{S(T_1, T_2)}{f(L(T_1))f(L(T_2))}$$

Für eine solche Operation hat sich der Begriff *Normierung* eingebürgert, wobei f in dieser Beschreibung als *Norm* von T, geschrieben als $\|T\|$, interpretiert würde.

Eine Parallele wäre das euklidische Skalarprodukt

$$\vec{x} \cdot \vec{y} = \|\vec{x}\|\|\vec{y}\| \cos(\alpha) \quad ,$$

wobei α den Winkel zwischen den beiden Vektoren $\vec{x}$ und $\vec{y}$ bezeichnet. Teilt man das Skalarprodukt durch die Normen (Längen) beider Vektoren, so bleibt nur noch der Kosinus übrig, der von der relativen Orientierung beider Vektoren abhängig ist. Wenn man nun nur an der relativen Orientierung von Vektoren interessiert ist, und nicht an ihrer Länge, hat man mit dem Kosinus ein brauchbares Maß für die Ähnlichkeit der Richtung gefunden. Unter dem Eindruck dieser Parallele könnte man hoffen, mit einer einfachen Operation die für jeweils nur von einem Text abhängigen Anteile abzuspalten. Dies ist letztlich genau das Verfahren, das im Bereich des *Information Retrieval* als „Vektorraummodell“[30] etabliert ist (Salton et al., 1975).

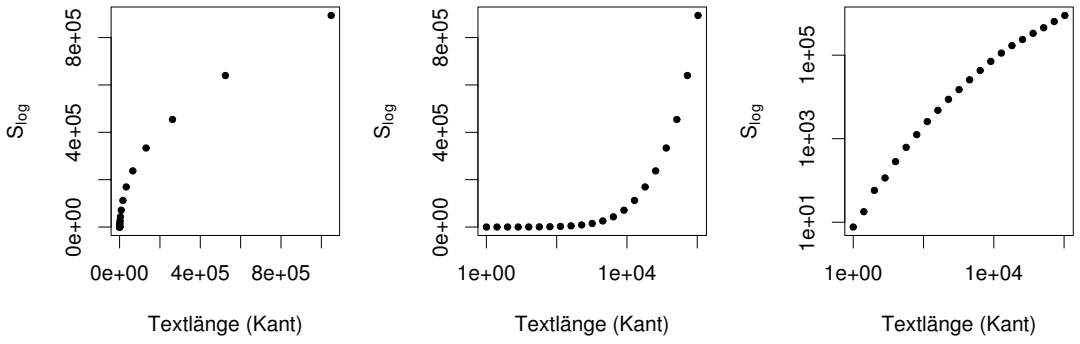
[30] Mathematisch ist diese Metapher etwas weit hergeholt, da Texte schwer als Vektoren im strengen Sinn interpretiert werden können, da zum Beispiel der Begriff eines „negativen“ Textes fehlt.

Die genaue Form dieser Textlängenabhängigkeit ist allerdings nicht ohne weiteres zu modellieren oder nur unter Zuhilfenahme von möglicherweise unrealistischen Modellannahmen. So ließe sich wahrscheinlich für randomisierten Text eine explizite Abhängigkeit von der Textlänge analytisch berechnen, damit hätte man aber nicht unbedingt etwas für den Fall natürlicher Sprache gelernt.

Um die Textabhängigkeit von S näher zu beleuchten greife ich das empirisch interessanteste Maß S_{log} heraus und variiere die Länge eines der eingehenden Texte.

Datengrundlage waren die zwei deutschen Texten Kant (2004) (*Kritik der reinen Vernunft*) und Fontane (2004) (*Effi Briest*). Aus dem Fontane-Text wurde jeweils die ersten $2^{19} \approx 520000$ Zeichen verwendet. Dieses sei mit $T_{fontane}$ bezeichnet.

Als zweiter Testtext wurden nacheinander die ersten 2^i Zeichen von Kant (2004) verwendet, wobei i von 0 bis 20 lief (2^{20} ist genau ein Mb Text). Dieses Teilstück sei T_{kant}. Mit Hilfe einfacher graphischer Darstellungen lässt sich überprüfen, ob die Textlängenabhängigkeit sich über eine lineare Funktion beschteiben lässt ($f(L(T)) = aL(T)$), in logarithmischer ($f(L(T)) = a\log(bL(T))$) bzw. eponentieller ($f(L(T)) = a\exp(bL(T))$) Form darstellbar ist, oder über ein Potenzgesetz beschrieben werden kann ($f(L(T)) = aL(T)^b$).[31]



(a) lineare Darstellung (b) logarithmisch in x (c) logarithmisch in x und y

Abbildung 3.1: $S_{log}(T_{fontane}, T_{kant})$ in Abhängigkeit von der Länge des Textes T_{kant}.

Abbildung 3.1 zeigt $S_{log}(T_{fontane}, T_{kant})$ aufgetragen über der Textlänge von T_{kant} in drei verschiedenen Darstellungsformen. Teilbild 3.1a zeigt beide Achsen in linearer Darstellung. Die Kurve steigt erst steil an und wird dann flacher. Lineares Verhalten ist damit wiederlegt, dieses Aussehen ist aber sowohl mit einer logarithmischen Abhängigkeit verträglich, als auch mit einem Potenzgesetz (z.B. mit der Wurzel der Textlänge, $f(L(T)) = aL(T)^{\frac{1}{2}} = a\sqrt{L(T)}$).

Teilbild 3.1b zeigt dieselben Daten mit logarithmischer x-Achse. Eine logarithmische Abhängigkeit müsste sich hier als eine gerade Linie zeigen. Dies ist offensichtlich nicht der Fall.

[31] Der lineare Fall ist mathematisch nur ein Spezialfall eines solchen Potenzgesetzes mit $b = 1$.

Entsprechend zeigt Teilbild 3.1c die Daten in doppellogarithmischer Darstellung. Jegliche x^b-Abhängigkeit, also auch eine Wurzel, müsste sich hier wiederum als eine gerade Linie zeigen. Auch dies findet sich nicht.

Für die anderen Maße gelten ähnliche Verhältnisse. Nur das lineare $S_l(T_{fontane}, T_{kant})$ lässt sich über einen gewissen Bereich als $S_l(T_{fontane}, T_{kant}) = aL(T_{kant})$ parametrisieren. Aber auch dies gilt nur eingeschränkt und ist nicht besonders hilfreich.

Das heißt, die Abhängigkeit der S-Maße, bzw. ihrer logarithmisch aufsummierten Anteile ist nicht leicht über eine polynomiale oder logarithmische Funktion zu parametrisieren. Man könnte versuchen, ein komplizierteres Modell zu aufzustellen oder eine rein empirisch bestimmte Form ansetzen, um die Abhängigkeiten von der Textlänge doch noch abzuspalten.

Es soll an dieser Stelle kurz diskutiert werden, ob dies überhaupt von praktischem Nutzen sein kann. Dafür müsste idealerweise die Textlängenabhängigkeit der einzige jeweils nur von einem Text abhängige Anteil sein. Das heißt, bei gleichlangen Texten sollte es von Anfang an keine *einzel*textspezifischen Anteile geben. Dies wird im Folgenden überprüft und widerlegt. Abbildung 3.2 zeigt eine Darstellung des Korpus, das

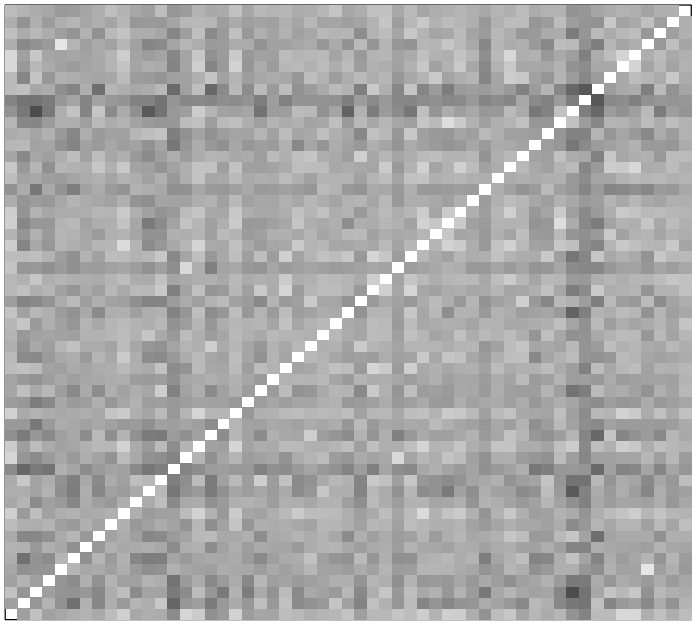


Abbildung 3.2: 55 gleich lange Texte im Vergleich. Die Reihen und Spalten der Matrix stehen jeweils für die Texte T_i und T_j. Die Felder kodieren die $S(T_i, T_j)$-Werte. Hohe Werte korrespondieren mit hellen Feldern.

in Abschnitt 3.6.4 noch eingehend Thema sein wird. Dieses Korpus enthält 55 unter kontrollierten Bedingungen erhobene Aufsätze 17-jähriger Jugendlicher.

Alle Dateien sind hier reduziert auf die Länge des kürzesten Textes. Jedes Feld der Matrix entspricht einem Wert $S_{log}(T_i, T_j)$, wobei die T_i (Zeilen) und T_j (Spalten) die erwähnten 55 gleichlangen Texte durchlaufen. Hellere Felder entsprechen größeren S-Werten, dunklere den kleineren. Die Diagonale ist ausgespart, da die Werte für $S(T_i, T_j)$ mit $i = j$ undarstellbar groß und bedeutungslos sind.

Deutlich ist die symmetrische Form der Matrix und ein gewisses Streifenmuster zu erkennen. Dieses Streifenmuster besagt, dass zum Beispiel der neunte Text von oben, der einen erkennbaren dunklen Streifen bildet, mit den anderen Texten im Mittel ein ziemlich niedriges S bildet.

Das heißt, es scheint einen textlängenunabhängigen Anteil in S_{log} zu geben, der dennoch spezifisch für einen bestimmten Text ist. Das Streifenmuster in Abbildung 3.2 ist signifikant. Um dies zu überprüfen habe ich die Mittelwerte der horizontalen Streifen mit einer ANOVA miteinander verglichen. Jeder Streifen entspricht den möglichen $S_{log}(T_i, T_j)$ Werten für ein festes i. Um Komplikationen wegen der Symmetrie der Matrix zu vermeiden, habe ich mich dabei auf den linken oberen Quadranten beschränkt, so dass alle eingehenden $S_{log}(T_i, T_j)$-Werte unterschiedlich sind. Dass die vertikal untereinanderliegenden Matrixfelder sich einen Text T_j teilen, wurde als Messwiederholung berücksichtigt. Es ergibt sich ein p-Wert $< 2^{-16} \approx 0$. Diese Eigenschaft von S-Verteilungen ist einer der Punkte, an denen sich andeutet, dass die hier untersuchten Häufigkeitdaten Einblicke in die statistischen Eigenschaften geschriebener Sprache ermöglichen könnten.

Diese Anteile in S_{log} und genauso in allen anderen S-Varianten verhindern, dass selbst die radikalste Form der „Normierung“ Erfolg haben kann, nämlich alle Texte auf dieselbe Länge zurechtzustutzen.

Daher habe ich eine simplere, heuristische Methode der Normierung gewählt, die es vermeidet, die einzeltextspezifischen Anteile von S zu modellieren und den zusätzlichen Vorteil hat, sämtliche solche Anteile auf einmal zu nivellieren.

Das Vorgehen sei an einem anderen Datensatz erklärt, der in Abschnitt 3.5 noch Thema sein wird. In Abbildung 3.3 wird jeweils ein Text t_i (Reihen) mit einem Text T_j (Spalten) verglichen. Die Texte sind Essays niederländischer Studenten. Die T_i bestehen aus der Aneinanderreihung 5 solcher Texte die t_j enthalten nur einen. In der Matrix dargestellt ist $S_{log}(t_i, T_j)$. Wieder stehen dunkle Farben für kleine Werte von S_{log}.

Der Autor von T_i ist nun in jedem Fall auch der Autor von t_i. Das heißt, die Dateien gleicher Autorschaft liegen auf der Diagonalen. Die weißen Quadrate markieren nun für jede Datei t_j (Spalten) das Feld mit maximalem $S_{log}(T_i, T_j)$. Würde man nun auf diese Weise den Autor des entsprechenden Textes zuordnen wollen, läge man zwar in 5 von 8 Fällen richtig, in den übrigen 3 Fällen würde aber Autor #3 fälschlich als Autor ausgewählt.

Die Vermutung liegt nahe, dass das am besprochenen Textlängeneffekt liegt, das heißt, dass Datei T_3 wohl am längsten ist. Dies ist nicht der Fall, T_2 ist mit 28133 Zeichen etwas länger als T_3 mit 27732 Zeichen. Woran das liegt ist unklar, sei es dass T_3 besonders typisch für den allgemeinen Sprachgebrauch ist, sei es, dass es eine andere tiefer liegendere oder trivialere Erklärung gibt. Dies mag als Fragestellung für weitere Forschung beste-

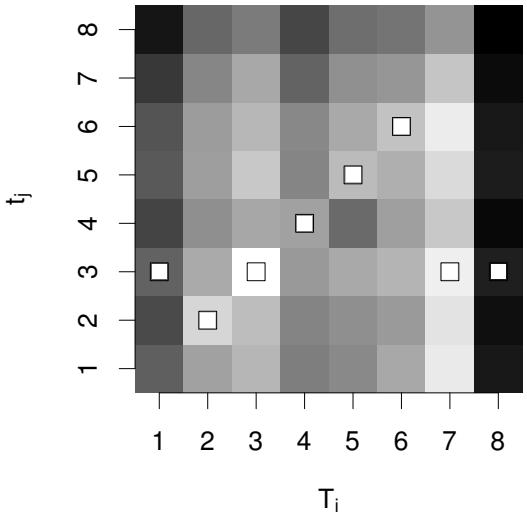


Abbildung 3.3: Dateien aus Juolas Testkorpus (Problem M) im Vergleich. Dargestellt sind die $S_{log}(t_i, T_j)$-Werte für Testtexte t_i und längere Trainingstexte T_j. Hellere Graustufen entsprechen höheren S_{log}-Werten. Weitere Details im Text.

hen bleiben. Hier geht es im Folgenden um die Entwicklung einer heuristischen aber effektiven Methode, die einzeltextspezifischen Anteile aus S herauszurechnen.

Man macht sich leicht klar, dass die sehr deutlichen vertikalen Streifen verschwinden, wenn man die Spalten jeweils durch ihren Mittelwert teilt. Dasselbe gilt für die horizontalen Streifen, die zwar schwerer zu erkennen, aber für unsere Fehlzuweisungen verantwortlich sind. Das Ergebnis einer solchen heuristischen Normierung ist in Abbildung 3.4 dargestellt. Nun werden 7 von 8 Dateien den richtigen Autoren zugeordnet, eine wesentliche Verbesserung.

In mathematischer Notation:

$$S_{norm}(t_i, T_j) = \frac{S(t_i, T_j)}{\frac{1}{n}\sum_{j'=1}^{n} S(t_i, T_{j'}) \frac{1}{n}\sum_{i'=1}^{n} S(t_{i'}, T_j)} \tag{3.2}$$

Hier steht S und auch S_{norm} für jede der in Abschnitt 3.3 definierten Varianten.

Dies ist nicht die einzige denkbare Form der Normierung, sondern lediglich eine einfache Heuristik, die sich in der Praxis als äußerst effektiv erwiesen hat. in Abschnitt 3.6.4 wird für einen Spezialfall eine analytisch besser begründete Version der Normierung untersucht werden. Das vorweggenommene Ergebnis dort ist, dass sich damit zwar Performanzvorteile ergeben, diese aber vergleichsweise gering sind.

In einem wesentlichen Punkt ist das hier verwendete Verfahren einzigartig in der mir bekannten Forschungsliteratur: Durch die Normierung wird die verfügbare Information von allen im Korpus enthaltenden Dokumenten miteinander verbunden. Dies ist genau

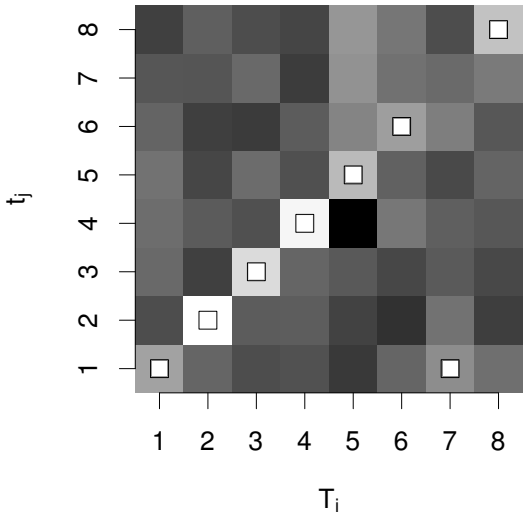


Abbildung 3.4: Dateien aus Juolas Testkorpus (Problem M) im Vergleich. Dargestellt sind die $S_{log}(t_i, T_j)$-Werte für Testtexte t_i und längere Trainingstexte T_j. Hellere Graustufen entsprechen höheren S_{log}-Werten. Im Gegensatz zu Abbildung 3.3 sind Zeilen und Spalten gemäß Gleichung 3.2 normiert.

das, was die Überlegenheit der maschinellen Lernverfahren in der Literatur begründen könnte (Vergleiche Seite 142).

Ein weiterer Unterschied verdient Erwähnung. Im Gegensatz zur Mehrzahl der in Abschnitt 3.2 dargestellten Maschinenlernverfahren enthält das hier vorgestellte Klassifikationsverfahren keine freien numerischen Parameter, die einer Anpassung bedürften. Die endgültige Klassifikation folgt direkt und eindeutig aus den vollständigen Substringhäufigkeiten der verglichenen Texte. Zur Rolle von *tuning parameters* in stilometrisch eingesetzten Maschinenlernverfahren vergleiche auch Jockers und Witten (2010).

3.5 Ein empirischer Vergleich der definierten Ähnlichkeitsmaße

In diesem Abschnitt soll die Frage geklärt werden, welches der in Abschnitt 3.3 vorgestellten Ähnlichkeitsmaße am effektivsten in der Lage ist, Texte desselben Autors von Texten verschiedener Autoren zu unterscheiden.

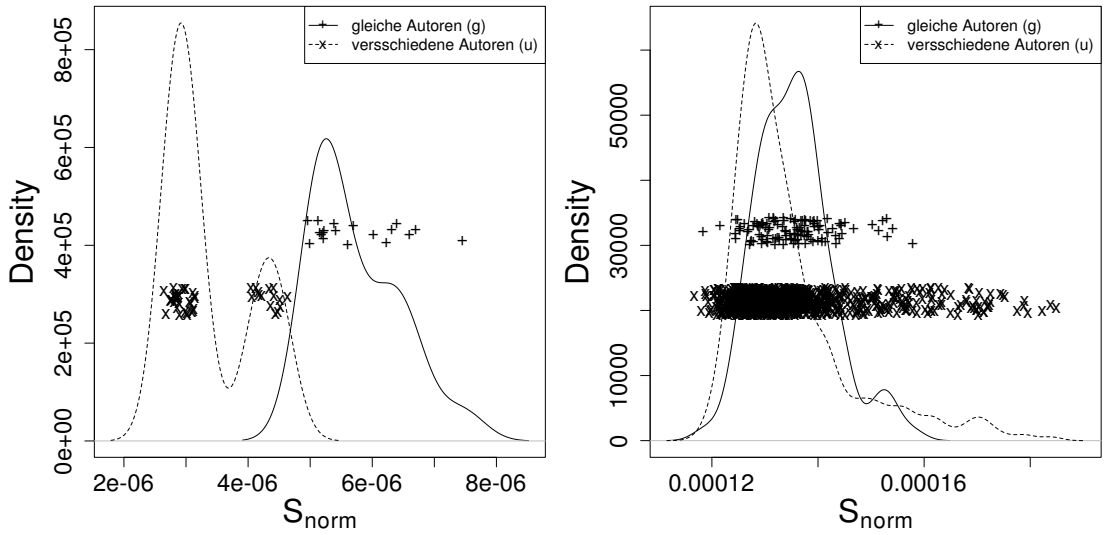
Zu diesem Zweck evaluiere ich ihre Performanz auf dem für derartige Tests entworfenen Datensatz[32] von Juola (2004) . Dieses Korpus enthält 8 verschiedene Subkorpora, die jeweils verschieden gelagerte Probleme zur Autorschaftsbestimmung darstellen. Jedes dieser Subkorpora besteht aus einer unterschiedlichen Anzahl Test- und Trainingstexte.

[32] Das Korpus und vorläufige Resultate des durchgeführten Wettbewerbs können unter `http://www.mathcs.duq.edu/~juola/authorship_contest.html` (besucht am 18.10.2012) heruntergelanden werden.

Die ursprüngliche Fragestellung bestand darin, die Testtexte möglichst sicher ihren Autoren zuzuordnen. Leider ist es mir nicht gelungen, die tatsächliche Zuordnung der Testtexte zu ihren Autoren in allen Fällen in Erfahrung zu bringen. Daher arbeite ich hier ausschließlich mit den Trainingstexten und ziehe diese auch als Testtexte heran (natürlich nicht gleichzeitig).

Vergleicht man alle n Texte eines Subkorpus untereinander und normiert anschließend wie in Abschnitt 3.3 beschrieben, ergeben sich $n(n-1)$ normierte S-Werte, wenn die Diagonale ausgespart bleibt. Für die symmetrischen Maße S_{log}, S_{mlog} und S_{shlog} sind nur die Hälfte dieser Werte unterschiedlich, so dass wir von hier an nur mit diesen $n(n-1)/2$ Werten arbeiten werden.

Diese lassen sich in zwei Klassen einteilen: In der Teilmenge g sind alle S_{norm}-Werte enthalten, die Texte desselben Autors miteinander vergleichen. u enthält alle übrigen S_{norm}-Werte. Für zwei beispielhafte Teildatensätze ist die Verteilung von g und u in Abbildung 3.5 miteinander verglichen.



(a) Teilkorpus D

(b) Teilkorpus M

Abbildung 3.5: Verteilung der S_{norm}-Werte für den Fall, dass Texte gleicher (g), bzw. verschiedener (u) Autoren verglichen werden. Beide Teilbilder stellen die *logarithmische* Variante S_{log} dar. Die einzelnen Textvergleiche sind als `+` (für g) und `x` (für u) eingezeichnet. Die Y-Koordinate ist willkürlich. Zusätzlich zeigen die durchgehenden und gestrichelten Linien die Verteilung von g bzw. u genähert durch eine kontinuierliche Kurve. Für Subkorpus M wird so erst der entscheidende Mittelwertsunterschied zwischen g und u sichtbar. Bemerkenswert ist die kleine Varianz von S_{norm}: Sie liegt in Teilbild (b) in der Größenordnung von 10% des Erwartungswertes. In vielen Fällen ist das Verhältnis noch wesentlich kleiner. Siehe dazu Golcher (2007a).

Man sieht, dass für die zwei dargestellten Subkorpora Vergleiche von Texten gleicher Autoren im Mittel zu einem höheren S_{norm} führen. Im Bild sind lediglich die Werteverteilungen zu sehen, die sich für den *logarithmischen Ähnlichkeitsindex* S_{log} ergeben. Die Graphik ist ein erster Hinweis daraus, dass sich mit den S-Maßen tatsächlich Stilometrie betreiben lässt.

Doch welche der definierten Varianten ist am effektivsten in der Lage, Texte gleicher Autoren zu identifizieren? Informell lässt sich das als die Frage formulieren, welches S jeweils die beiden Punktewolken in Abbildung 3.5 am klarsten von einander trennt.

Es gibt verschiedene Tests, um zu überprüfen, ob die Mittelwerte zweier Stichproben signifikant unterschiedlich sind. Die Teststatistiken dieser Tests[33] können entsprechend als ein Maß für die Unterschiedlichkeit der beiden Stichproben betrachtet werden. Es ist zu beachten, dass diese Überlegung nichts mit der Signifikanz des Mittelwertunterschieds zu tun hat, die Aussagen über die zugrundeliegenden Grundgesamtheiten erlaubt. Hier soll es erst einmal nur darum gehen, den Unterschied von g und u zu quantifizieren und in Bezug auf die zugrundeliegende Variante von S zu vergleichen.

Für die erwähnten Signifikanztests werden gemeinhin der t-Test oder der *Wilcoxon-Rangsummentest* (Bortz, 2005, S.140 bzw. S.153) verwendet. Ersterer sollte nur herangezogen werden, wenn die Daten halbwegs normalverteilt sind, was zum Beispiel für das Subkorpus D (Teilbild 3.5a) ganz offensichtlich nicht gegeben ist. Obwohl wir zwar nicht direkt an einem gültigen Signifikanztest interessiert sind, wird daher zur besseren Interpretierbarkeit der Ergebnisse nicht das Student'sche t, sondern Wilcoxons W herangezogen. Die Ergebnisse sind in Tabelle 3.1 dargestellt.

	A	C	D	F	G	I	K	M
S_{log} (Def. 34)	15772	2450	893	586115	16	21	1924	73759
S_{mlog} (Def. 35)	15511	2443	893	576623	16	20	1923	73027
S_{shlog} (Def. 37)	14261	2424	866	522521	25	15	218	67660
$S_{llog/rlog}$ (Def. 36)	14307	2430	893	539965	25	15	1599	72848
S_l (Def. 33)	12191	2199	694	476226	27	16	386	62684

Tabelle 3.1: Die definierten S-Maße im Vergleich. Gezeigt ist für die Teilkorpora von Juolas Testkorpus das Wilcoxon'sche W für den Vergleich von S-Werten für gleiche und für verschiedene Autoren. Die Werte für S_{llog} und S_{rlog} sind gemeinsam gezeigt, da sie dieselbe Performanz haben müssen, wenn jeder Text einmal als Testtext und einmal als Trainingstext auftritt.

Die Werte der Tabelle erlauben einen genauen Vergleich und auch eine Abschätzung der relativen Bedeutsamkeit der Ergebnisse. So ist Abteilung I mit nur 5 Dokumenten das kleinste Unterkorpus, während Abteilung F mit 68 Dateien bedeutend größer ist. Um die Daten noch einmal in übersichtlicherem Format darzustellen sind in Abbildung 3.6 die sich aus der Tabelle ergebenden Rangplätze graphisch dargestellt.

In allen Fällen bis auf Subkorpus G schneidet die *logarithmische* Version S_{log} am besten ab, für die beiden Subkorpora D und G liegen S_{log} und der *symmetrisch halblogarithmis-*

[33] z.B. der t-Wert des t-Tests

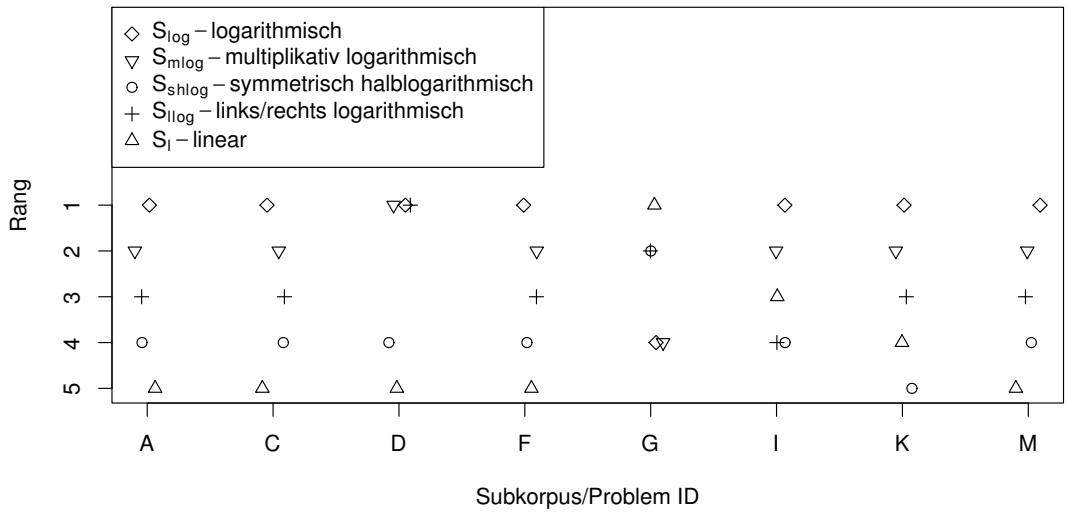


Abbildung 3.6: Visualisierung der Rangplatzierungen der verschiedenen Maße, die sich aus Tabelle 3.1 ergeben. Je größer das dort angegebene W, desto besser können Textpaare mit gleichem Autor von den übrigen Textpaaren getrennt werden. Entsprechend entspricht die Performanzreihenfolge der Maße für Problem A der Anordnung der Tabellenzeilen: $S_{log} > S_{mlog} > S_{shlog} > S_{llog/rlog} > S_l$. Diese Reihenfolge wird von der Reihenfolge der Symbole in der Graphik dargestellt. In 7 von 8 Teilkorpora liegt S_{log} an erster Stelle. Auch die übrigen Maße verhalten sich jeweils sehr ähnlich. Nur für Problem G fgilt im Wesentlichen die umgekehrte Reihenfolge $S_l > S_{llog/rlog} = S_{shlog} > S_{log} = S_{mlog}$. Jeweils zwei Maße teilen sich hier einen Rangplatz.

che Ähnlichkeitsindex gleich auf. Das Subkorpus G stellt auch von seiner Datenbasis her eine Ausnahme dar, da es der einzige Datensatz ist, der nur Texte von einem einzigen Autor enthält, Edgar Rice Burroughs[34]. Diese sind unterteilt in Texte des „älteren“ und des „ jüngeren“ Burroughs (Juola, 2004) . Man muss nicht besonders beunruhigt sein, dass der vorgestellt Algorithmus nicht in der Lage ist, die verschiedenen Schaffensperioden desselben Autors zu unterscheiden. Im Authorship-Attribution-Contest, für den der Datensatz entwickelt wurde gelang es den 13 Teilnehmern im Mittel nur 1.8 von 4 Testdokumente der richtigen Schaffensperiode zuzuordnen, bei einer Baseline von 2 (Juola, 2004). Es gibt also keinen Hinweis darauf, dass sich das Alter von Burroughs überhaupt messbar auf seinen Schreibstil ausgewirkt hat. Wir vernachlässigen das Subkorpus G also für die folgenden Überlegungen.

Vom ersten Eindruck her scheint folgende Reihenfolge absteigender Performanz zu

[34] 1875–1950, amerikanischer Autor, geistiger Vater von „Tarzan“.

bestehen

$$S_{log} \quad \text{besser als} \quad S_{mlog} \quad \text{besser als} \quad S_{llog/rlog} \quad \text{besser als} \quad S_{shlog} \quad \text{besser als} \quad S_l$$

Dass S_{log} und S_{mlog} signifikant besser abschneiden als das *lineare* S_l und das *symmetrisch halblogarithmische* S_{shlog} scheint offensichtlich und ist leicht zu zeigen, selbst wenn man Subkorpus G außen vor lässt. Betrachten wir der Einfachheit halber nur S_{mlog} und S_{shlog}. Ausgangspunkt ist die Nullhypothese, dass beide Maße gleich gut sind. S_{log} schneidet in 7 der verbliebenen 7 Fälle besser ab. Ein einfacher Binomialtest ergibt einen p-Wert von 1.5%. Damit kann man die Hypothese verwerfen, dass die Überlegenheit von S_{log} und S_{mlog} zufällig ist.

Etwas komplizierter wird es, wenn man S_{log} und S_{mlog} miteinander vergleichen möchte. Hier ergibt sich für Subkorpus D für beide Maße ein identischer W-Wert. Dieser Wert ist im Rahmen des Binomialtests nicht deutbar.

Daher ist man gewzungen, direkt auf die Werte S_{log} und S_{mlog} in den einzelnen Subkorpora zurückzugreifen. Für jeden der S_{log}-Werte wird der relative Abstand vom Mittelwert aller Werte $r_{log} = {}^{S_{log}}/_{\overline{S_{log}}}$ bestimmt. Das geschieht für jedes Unterkorpus getrennt.

Dieselbe Renormierung geschieht nun mit den $r_{mlog} = S_{mlog}/\overline{S_{mlog}}$-Werten. Nun gibt es zwei lange Vektoren mit r_{log} und r_{mlog}-Werten.

In einem letzten Schritt gilt es nachzuweisen, dass das Verhältnis r_{log}/r_{mlog} größer ist als 1, falls beide Texte vom selben Autor sind, während es sonst kleiner ist. Wenn diese Hypothese stimmt, so wäre r_{log} besser geeignet, die zwei Fälle voneinander zu trennen, da die S_{log}-Werte jeweils in der richtigen Richtung weiter vom Mittelwert entfernt sind. Ich fasse beide Fälle in einer einzigen Definition zusammen:

$$d_{log/mlog} = \begin{cases} r_{log}/r_{mlog} & \text{falls die verglichenen Autoren identisch sind} \\ r_{mlog}/r_{log} & \text{sonst} \end{cases}$$

Mit anderen Worten: $d_{log/mlog}$ ist ein Maß für die relative Auflösung von S_{log} und S_{mlog}. Wenn d im Mittel größer ist als 1, so sind die S_{log}-Werte im Mittel weiter von ihrem Mittelwert entfernt als die S_{log}-Werte. Je weiter diese Maße von ihrem Mittelwert entfernt sind, desto klarer unterteilen sich die Verteilungen in die beiden Gruppen der Textpaare mit identischem oder unterschiedlichem Autor.

Abbildung 3.7 zeigt beispielhaft die Verteilung von $d_{log/mlog}$ für das Subkorpus M. Die Graphik lässt bereits vermuten, dass die Verteilung signifikant von einem Mittelwert 1 abweicht. Dies bestätigt wiederum der *Mann-Withney-Test*: Für die dargestellten Daten ergibt sich ein p von $1.5 \cdot 10^{-12}$.

Auch für einen simultanen Vergleich aller $d_{log/mlog}$-Werte mit 1 ergeben sich p-Werte von sehr nahe Null.

Mit derselben Klarheit lassen sich alle anderen Glieder der Ungleichung 3.5 überprüfen. Am geringsten ist insgesamt der Abstand zwischen S_{shlog} und $S_{llog/rlog}$.

Es ist festzuhalten, dass die Maße, die in beiden Texten logarithmisch sind, denjenigen Maßen, die nur in einem Text oder gar nicht logarithmisch sind, überlegen sind. Die nichtlogarithmischen Maße sind aber genau diejenigen, die sich am direktesten auf das im Vek-

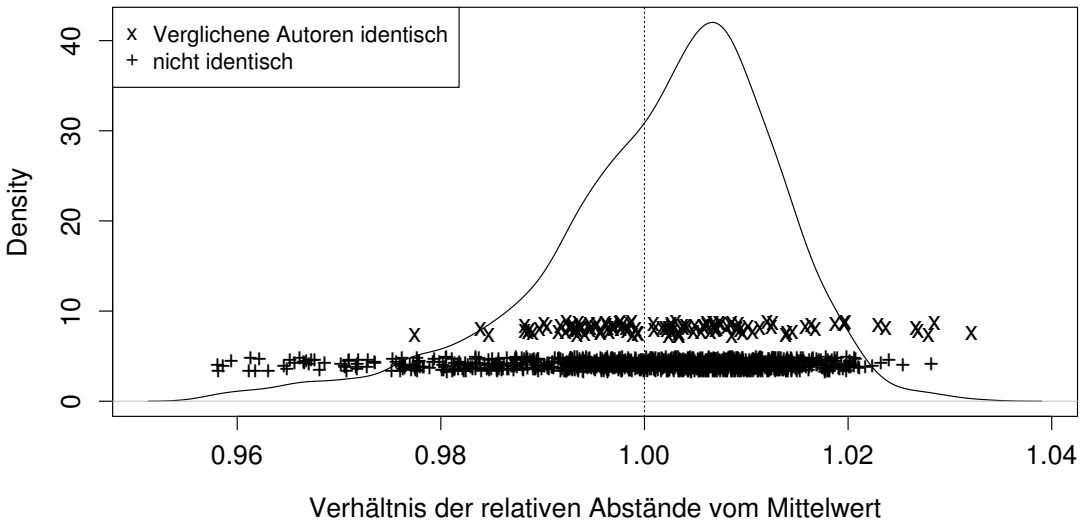


Abbildung 3.7: Die Verteilung von $d_{log/mlog}$ für das Subkorpus M. Zur Illustration sind die einzelnen Datenpunkte ebenfalls eingetragen, getrennt danach, ob die Texte identischer Autoren verglichen wurden, oder nicht. Es wurde auch ein Test durchgeführt, ob beide Punktemengen sich in ihrem Mittelwert unterscheiden (*Mann-Withney-Test*). Dies ist nicht der Fall ($p = 0.66$).

torraummodell übliche Skalarprodukt abbilden lassen. Wie im Forschungsüberblick 3.2 dargelegt gibt es in der Stilometrie insgesamt eine sehr starke Tendenz, für die Analyse lediglich die häufigsten und/oder kürzesten n-Gramme heranzuziehen. Gerade in diesem Bereich ist der Effekt des Logarithmus nicht besonders groß, da dieser erst wirklich zum Tragen kommt, wenn Zahlen sehr unterschiedlicher Größenordnungen verglichen werden. Entsprechend ist die Verwendung von logarithmischen Frequenzdaten eine Ausnahme. Eindrücklich zusammengefasst ist das „lineare“ Paradigma in Forsyth et al. (1999, 6), wo es unter dem Namen „Burrows Approach“ beschrieben wird. Die in nur einem Text logarithmischen Maße wiederum entsprechen in ähnlicher Weise den in Arbeiten wie Teahan (2000) verwendeten entropiebasierten Maßen (s. Abschnitt 3.3, Seite 148).

Die hier gefundene Performanzreihenfolge der S-Maße wird im folgenden anhand variierter Fragestellungen und in unterschiedlichen Daten systematisch bestätigt werden können. Es ist eine bemerkenswerte und interessante Beobachtung, dass wir nun zwei sehr unterschiedliche Fragestellungen kennen, die mit Hilfe identischer Daten bearbeitet wurden und zum selben Schluss führen: Sowohl in der Morphologischen Induktion als auch in der Stilometrie erweist sich die logarithmierte Form der Frequenzdaten als die überlegene Darstellung. Aus linguistischer Sicht kann man aus diesen Ergebnissen den Schluss ziehen, dass in den längeren und selteneren Zeichenketten durchaus genug Information steckt um einen konsistent messbaren Performanzunterschied hervorzurufen. Es scheint möglich, dass dieser überraschende Informationsgehalt längerer Zeichenketten mit den in der Einleitung 1 referierten langreichweitigen Korrelationen in Texten zusammenhängt.

In der mir bekannten Literatur wird zum Kontrast der linearen und logarithmischen Repräsentation von Frequenzdaten selten Stellung genommen. Eine Ausnahme ist Diederich et al. (2003). Die Autoren beschäftigen sich mit *Authorship Attribution* mit Hilfe von SVM. Sie kommen in ihrer Untersuchung zu Ergebnissen, die meine qualitativ sehr ähnlich sind:

> [...] logarithmic relative frequencies with L1 normalization have the overall best performance. (Diederich et al., 2003, 118)
>
> [...] Simple relative frequencies do much worse than the logarithmic version. (Diederich et al., 2003, 120)

„$L1$ normalization“ ist eine Alternative zur bekannteren Euklidischen Distanz, die – wie der Logarithmus – den häufigeren Elementen des Frequenzvektors weniger Gewicht gibt. Leider ziehen die Autoren keine weitergehenden Schlüsse aus ihren Beobachtungen.

Da sich S_{log} als optimal erwiesen hat, wird im Folgenden das Hauptaugenmerk auf diesem Maß liegen.

Die Ergebnisse des *Ad-hoc Authorship Attribution Contests* (Juola, 2004) wurden veröffentlicht (Juola, 2006a). [35]. Dies erlaubt es, die damals verwendeten Ansätze mit der Performanz meines eigenen zu vergleichen. Welche das sind ist am detailliertesten in Juola (2006a, 292) beschrieben. Evaluationsmaß ist der Anteil der korrekt klassifizierten Texte, die *Accuracy*. Die Ergebnisse sind in Abbildung 3.8 dargestellt. Wie oben erwähnt, kenne ich die Auflösung nicht. Dies erlaubt es mir nur mit den Trainingsfiles zu arbeiten. Das heißt, meine Trainingskorpora waren kleiner und die klassifizierten Dokumente waren andere. Die Verkleinerung des Trainingskorpus sollte die Performanz eher herabsetzen, während die unterschiedlichen Testtexte keinen systematischen Effekt haben sollten. Dies macht es wahrscheinlich, dass die Methode auch bei der Klassifikation der tatsächlichen Testtexte den Vergleichsmethoden in der überwiegenden Mehrzahl der Teilprobleme überlegen wäre.

3.6 Untersuchungen zu verschiedenen stilometrischen Aufgabenstellungen

Im vorhergehenden Abschnitt wird die generelle Anwendbarkeit der Methode auf stilometrische Fragestellungen anhand eines extra zu Wettberbs- bzw. Testzwecken entwickelten Korpus untersucht. Auf dieser Datenbasis konnte auch eine eindeutige Performanzreihenfolge der in 3.3 definierten Varianten von S abgeleitet werden. Ein Testkorpus wird darauf hin entwickelt, über die eigentliche Fragestellung hinausgehende Probleme möglichst auszublenden.

Daher soll der Algorithmus nun an einer Reihe spezialisierter Fragestellungen untersucht werden. Drei davon waren bereits Thema stilometrischer Untersuchungen. Dies erlaubt eine gewisse Vergleichbarkeit mit veröffentlichen Performanzwerten herzustellen.

[35] Meine Gegenüberstellung basiert auf den vorab online veröffentlichten Version der Ergebnisse (Juola, 2004). Diese ist weitestehend identisch.

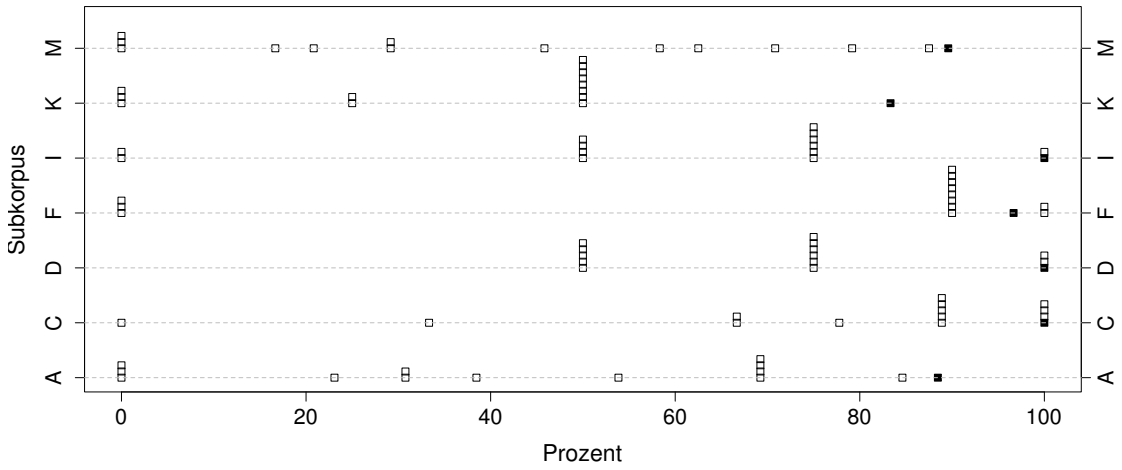


Abbildung 3.8: Vergleich meiner Methode mit den vorläufigen Ergebnissen (Juola (2004), weitgehend identisch mit Juola (2006a)). Die offenen Quadrate repräsentieren die Ergebnisse der Wettkampfteilnehmer, die gefüllten Quadrate meine eigenen Resultate. Problem G wurde nicht in die Übersicht aufgenommen, da es Zweifel an der Wohldefiniertheit der Fragestellung gibt. Vgl. Abbildung 3.6 und die Diskussion dazu auf Seite 3.5.

In Abschnitt 3.6.1 wird anhand eines von Baroni und Bernardini (2006) erstellten Korpus der Frage nachgegangen ob und wie gut sich übersetzte von nicht übersetzten Texten unterscheiden lassen. In Abschnitt 3.6.2 werden Texte aus dem ICLE (Granger, 2003) nach der Muttersprache ihrer Autoren klassifiziert. In 3.6.3 wird ein altes Standardproblem der Automatischen Autorenbestimmung untersucht, die *Federalist Papers*.

Die in Abschnitt 3.6.4 thematisierte Fragestellung ist meines Wissens noch nicht im Rahmen der Stilometrie untersucht worden: Ist Stil vererbbar? Dies wird anhand eines Korpus von Aufsätzen untersucht, deren Autoren sich aus Zwillingspaaren zusammensetzen.

3.6.1 Translationese: Eigenheiten übersetzter Texte

Beim Lesen übersetzter Texte meint man manchmal zu spüren, dass es sich um eine Übersetzung handelt. Sei es, dass die Wortwahl an eine bekannte fremde Sprache erinnert, sei es, dass eine Satzkonstruktion unvertraut anmutet.

Gellerstam (1986) prägte für dieses Phänomen den Begriff *Translationese*. In seiner heuristisch geprägten Arbeit suchte er nach signifikanten Unterschieden zwischen original schwedischen Texten und Übersetzungen aus dem Englischen ins Schwedische.

Seitdem wird daran geforscht, ob dieser „Dialekt der Übersetzung“ tatsächlich existiert und auf welcher sprachlichen Ebene und anhand welcher Eigenschaften er sich

genau manifestiert. Auf die Frage nach Struktur und Herkunft der Unterschiede zwischen Übersetzungen und Zielsprache gibt es zwei mögliche Antworten. Zum einen wurde argumentiert, dass in der Übersetzung Charakteristika der Quellsprache durchscheinen („source language shining-through“, Teich (2003)). Es ist in der Tat kaum vorstellbar, dass die Quellsprache, bzw. der originale Wortlaut des übersetzten Textes keinerlei Einfluss auf die Übersetzung hat. So weisen zum Beispiel Dai und Xiao (2011) einen solchen Effekt anhand der Häufigkeit von Passivkonstruktionen in englisch-chinesischen Übersetzungen nach.

Daneben gibt es aber Hypothesen, die ein überraschenderes Phänomen implizieren, nämlich dass *Translationese* tatsächlich als ein sowohl von der Zielsprache als auch von den verschiedenen Quellsprachen getrenntes sprachliches System existiert. Verschiedene *Universalien* wurden als allgemeine Eigenschaften von Übersetzungen vorgeschlagen (Baker, 1993, 1996; Toury, 1995) und empirisch untersucht (Laviosa, 2002, 1998; Amit et al., 1994).[36] Diese Arbeiten sind nicht stilometrischer Natur, ich möchte daher auf Details hier nicht weiter eingehen. Eine genauere Besprechung wäre auch deshalb nicht sehr hilfreich, da die S-basierte Klassifikation auf dem jetzigen Stand keine Erkenntnisse zur genauen Natur von *Übersetzungsuniversalien* beitragen könnte. Die Methode ist aber durchaus in der Lage die Hypothese einer von der Zielsprache zumindest teilweise unabhängigen Varietät *Translationese* mit empirischen Daten zu untermauern.

Aktuellere, in Teilen *stilometrische* Arbeiten zum Thema sind Borin und Prütz (2000); Teich (2003); Baroni und Bernardini (2006); Corpas et al. (2008); van Halteren (2008); Kurokawa et al. (2009); Shlesinger (2009); Ilisei et al. (2010); Dai und Xiao (2011). Vor allem Dai und Xiao (2011) bieten einen konzisen Überblick über die Debatte zu *translation universals* und *shining-through*.

Ich gehe hier näher auf die Arbeit von Baroni und Bernardini (2006) ein. Sie sind die ersten, die die Suche nach *Translationese* in eine stilometrische Fragestellung kleiden: Es gilt, Texte, die sich sonst in möglichst allen Parametern wie Textthema oder Domäne, Genre, Entstehungszeit und ähnlichem gleichen, danach zu klassifizieren, ob es sich um Übersetzungen handelt, oder nicht.

Die Autoren analysieren die subtilen regelmäßigen Abweichungen zwischen Übersetzungen ins Italienische und Texten, die auf italienisch verfasst wurden. Im folgenden werde ich Texte, die keine Übersetzungen sind, als *Originale* bezeichnen. Die Autoren untersuchen einen Datensatz aus 813 Artikeln eines italienischen geopolitischen Magazins (*limes*[37]). 569 dieser Artikel sind Originale, die übrigen 244 sind qualitativ hochwertige Übersetzungen aus verschiedenen Sprachen[38]. Baroni und Bernardini (2006) gelingt die

[36] Der Beitrag von Amit et al. (1994) ist aus der Reihe zu Arbeiten über die statistischen Eigenschaften von Texten, die schon in der Einleitung Kapitel 1 diskutiert wurden. Die Autoren untersuchen *langreichweitige Korrelationen* in der hebräischen Originalfassung der Bibel und in Übersetzungen in verschiedene Sprachen. Sie finden diese in der Originalfasssung am ausgeprägtesten. Ihre Schlussfolgerung lautet: „Any translation, even the most faithful, must break the special literary style which is unique to the author. In addition, since the translator is obligated to the informational content of the original, its own literary style cannot be utilized, and therefore we observe a reduction of the long-range correlations.“.

[37] `temi.republica.it/limes`

[38] u.a. Englisch, Arabisch, Französisch, Spanisch, Russisch.

Erkennung von Originalen und Übersetzungen auf ihrem Korpus mit einer *Accuracy* von 86.7%. Die Vielfalt der Quellsprachen soll sicherstellen, dass sich deren Einfluss weitgehend herausmitteln:

> [The] translations are carried out from several source languages into Italian; thus, any effect we find is less likely to be due to the „shining-through“ of a given language (Teich, 2003), than to a more general translation effect. (Bernardini und Baroni, 2006)

Seit Baroni und Bernardini ihre Ergebnisse vorgestellt haben, ist bewiesen, dass es mit maschinellen Lernverfahren möglich ist, zwischen Übersetzungen und Originalen zu unterscheiden. Damit ist ihnen ein objektiver Nachweis gelungen, dass es Eigenschaften geben muss, die die beiden Textarten unterscheiden. Dh., *Translationese* ist als objektiv messbares Phänomen etabliert.

Im Folgenden wird überprüft, wie gut die hier vorgestellte stilometrischen Methode diese Aufgabe zu lösen im Stande ist. Dazu ziehe ich einen Vergleich mit den von Baroni und Bernardini (2006) vorgestellten Ergebnissen.

Die Autoren haben mir ihr Korpus dankenswerterweise zur Verfügung gestellt. Dieses eignet sich besonders gut für einen ersten Test der Methode an einer echten stilometrischen Fragestellung, vor allem da weitere mögliche Einflussgrößen wie *Genre* oder *Topic*[39] sorgfältig konstant gehalten wurden. Diese Homogenität verhindert zwar die automatische Übertragbarkeit der Ergebnisse auf anders geartete Korpora, erlaubt aber andererseits eine isolierte Untersuchung einer einzigen stilometrischen Variable: Handelt es sich bei einem Text um eine Übersetzung, oder nicht?

Dies ermöglicht einen ersten Test der Methode an einer Fragestellung, auf den sie ohne weitere Modifikationen anwendbar ist. In einem späteren Kapitel 3.6.3 werde ich über diesen Standardfall hinausgehen.

Aber es soll hier nicht nur um einen direkten Vergleich mit einem modernen, etablierten maschinellen Lernverfahren. Über einen derartigen Beweis der grundlegenden Angemessenheit der Methode hinaus werden die folgenden Aspekte untersucht:

- Das Korpus existiert in mehreren Varianten. Neben der originalen Textform auch in verschiedenen flachen syntaktischen Annotationsebenen. Dies ermöglicht einen gewissen Einblick in den Beitrag, den diese Ebenen zum Gesamtphänomen *Translationese* leisten. Eine qualitative Diskussion relativ zu den von Baroni und Bernardini (2006) gewonnenen Erkenntnissen ist möglich.

- Baroni und Bernardini (2006) untersuchen Uni-, Bi-, und Trigramme. In der vorliegenden Untersuchung ermöglicht eine Randomisierung der Wortreihenfolge eine vergleichbare Untersuchung des Einflusses weiter reichender Textzusammenhänge.

- Was für eine Rolle spielt die Menge des Trainingsmaterials? Ab welcher Trainingstextlänge ist es möglich, übersetzte Texte zu erkennen?

[39] Etwas präziser ausgedrückt: Die Domäne bzw das „macro-topic“ (Baroni und Bernardini, 2006) ist konstant (Geopolitik) während die speziellen Themen der Texte sich gleichmäßig verteilen. S. dazu auch Fußnote 44.

- Wie verhalten sich die Varianten von S? Lässt sich die in Abschnitt 3.5 gewonnene Performanzreihenfolge bestätigen? Gibt es Wechselwirkungen mit der (Trainings-)Textlänge?

Vor einer vergleichenden Darstellung meiner eigenen Untersuchungen und Ergebnisse folgt nun eine genauere Beschreibung des von Baroni und Bernardini (2006) verwendeten Korpus und der von ihnen durchgeführten Experimente.

Die verschiedenen Repräsentationen des Korpus erlauben es, die Frage zu untersuchen, welche Einheiten auf welcher Ebene für die erfolgreiche Klassifikation verantwortlich sind. Im einzelnen untersuchen Baroni und Bernardini (2006):

`tok`[40] Der ursprüngliche Texte, lediglich Eigennamen und Zahlen sind durch Platzhalter ersetzt. Dadurch sollen Probleme vermieden werden, die daraus folgen würden, wenn vielleicht bestimmte Eigennamen eher in Übersetzungen auftauchen als andere. So könnte man vermuten, dass die original italienischen Texte mehr italienische Namen enthalten. Die Platzhalter sind nummeriert (also z.B. NPR1, NPR2, ...). Für denselben Ausdruck wird innerhalb eines Dokumentes jeweils derselbe Platzhalter verwendet.

`lemma` alle Oberflächenwortformen sind durch ihre Lemmata ersetzt.

`pos` alle Oberflächenwortformen sind durch ihre POS-Tags ersetzt.

`mix`[41] Die Funktionswörter verbleiben in ihren Oberflächenformen, während die inhaltstragenden Wörter von ihren POS-Tags ersetzt werden. Hierbei gelten die häufigsten Adverbien ebenfalls als Funktionswörter.

Die Dokumente werden in Vektoren überführt, deren Dimensionen die Frequenzen von Unigrammen, Bigrammen und Trigrammen der Token der verschiedenen Repräsentationen sind. Dies ergibt $3 \cdot 4 = 12$ mögliche Grundkombinationen.

Die Autoren verwenden nicht alle jeweils möglichen Uni-, Bi-, und Trigramme, sondern filtern nach Frequenz. Sie experimentieren auch mit *tf*idf*-Gewichtung[42]. Hier werden Elemente höher bewertet, die spezifischer für einzelne Dokumente sind. Berücksichtigt man diese Variation, so ergeben sich 24 Repräsentationen der Texte.

Die eigentliche Klassifikation wird mit Hilfe von *Support Vector Machines* (SVM) ausgeführt. SVM ist ein Verfahren des überwachten maschinellen Lernens. Eine Menge von vorklassifizierten Beispielen wird in einem mehrdimensionalen Raum dargestellt. In diesem Raum wird eine (Hyper)-Ebene berechnet, die die Kategorien möglichst klar trennt. Neue Objekte werden klassifiziert, je nachdem auf welche Seite der Ebene sie fallen. Für eine genauere Beschreibung siehe Schölkopf und Smola (2002).

Die Qualität der Klassifikation wird mit 16-facher *Cross Validation* gemessen: In jedem Durchlauf werden je 15 Texte aus der Menge der Originale und 15 aus der Menge

[40] Bei Baroni und Bernardini (2006) bezeichnet als *wordform*.
[41] Bei Baroni und Bernardini (2006) bezeichnet als *mixed*.
[42] *term frequency–inverse document frequency*, Salton und McGill (1983)

der Übersetzungen herausgegriffen. Diese 30 Texte bilden die Testmenge, während die übrigen als Trainingskorpus verwendet werden.

Die üblichen Performanzmaße *Accuracy*, *Precision*, *Recall* und *F-score* (s. Definition 26, 27 und die Erklärungen vor Definition 32) werden berechnet, wobei das Erkennen einer Übersetzung als Erfolgsereignis gezählt wird.

Werden die 12 Arten von Häufigkeitsvektoren einzeln klassifiziert, erzielen die Autoren F-Werte zwischen 0.008 und 0.715. Abbildung 3.9 stellt ihre Ergebnisse bildlich

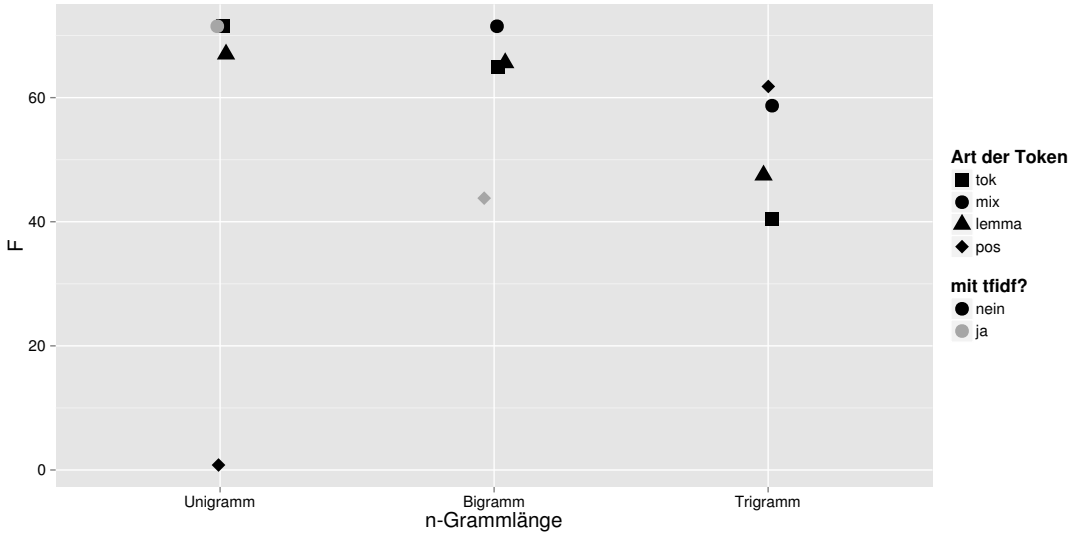


Abbildung 3.9: Visualisierung von Tabelle 4 aus Baroni und Bernardini (2006). Dargestellt ist das F-Measure der untersuchten Klassifikatoren auf einer Skala von 0 bis 100. Von links nach rechts nimmt die Kettenlänge zu. Man erkennt eine parallele Abwärtsbewegung der Repräsentationen, die lexikalische Information beinhalten (`tok`, `mix`, `lemma`), wenn man zu längeren Ketten übergeht. Dies ist plausibel mit der *Data Sparseness* in den Trigrammen dieser Repräsentationen zu erklären. Die `pos`-Repräsentation verhält sich gegenläufig.

dar. Die Repräsentationen, die lexikalische Information enthalten (`tok`, `lemma` und `mix`, wobei die letzte keine Inhaltswörter enthält) schneiden in der Trigramm-Variante deutlich schlechter ab als in den Uni- und Bigrammversionen. Für die `pos`-Repräsentation gilt das Gegenteil, hier nimmt die Performanz mit steigender Kettenlänge stark zu. Die Autoren führen dieses Verhalten überzeugend darauf zurück, dass in den lexikalisch informativen Repräsentationen die meisten Trigramme zu selten sind, um eine gute Klassifikation zu erlauben.

In einem zweiten Schritt kombinieren Baroni und Bernardini jeweils 4 bis 5 der bisher beschriebenen einfachen SVM-Klassifikatoren. Sie gehen dabei so vor, dass ein Text dann als Übersetzung klassifiziert wird, wenn mindestens eines der beteiligten Modelle ihn so einteilt. Dieses Vorgehen maximiert den *Recall*, der für die einfachen Modelle

wesentlich unterhalb der *Precision* liegt. Insgesamt untersuchen sie 24 unterschiedliche Kombinationen.

Die 10 besten dieser kombinierten Modelle schneiden mit F-Werten zwischen 0.844 und 0.862 deutlich besser ab als die einfachen Klassifikatoren. Das beste Modell kombiniert die Uni- und Bigramm-Modelle der `lemma`- und `mix`-Repräsentation und das Trigramm-Modell der `pos`-Repräsentation.

Ich halte mich in meiner eigenen Untersuchung möglichst exakt an das Vorgehen von Baroni und Bernardini. Allerdings verwende ich nicht die gleiche Menge an Repräsentationen der Texte. Die Repräsentationen `lemma` und `pos` standen mir nicht zur Verfügung. Neben den Repräsentationen `tok` und `mixed` habe ich noch weitere Varianten verwendet:

`txt` Der originale Text ohne die Ersetzung der Eigennamen wie in `tok`.

`fun` Ausschließlich die Funktionswörter der `mix`-Repräsentation.

`tag` Ausschließlich die POS-tags der `mix`-Repräsentationen.

`lex` Die Oberflächenwortformen zu den POS-tags in `tag`.

Alle so entstehenden Textrepräsentationen – bis auf `txt`[43] – habe ich wiederum in zwei Varianten untersucht:

`seq` Die Token in Originalreihenfolge.

`ran` Die Token in randomisierter Reihenfolge.

Vor der Besprechung der Klassifikationsresultate lohnt ein Blick auf die Verteilung der normierten S_{log}-Werte. Abbildung 3.10 visualisiert diese Verteilung. Berechnet wurde sie auf der `tok`-Repräsentation. Der Großteil der Daten ist annähernd normal verteilt. Die Textvergleiche, in die jeweils zwei Originale oder zwei Übersetzungen eingehen, liegen im Mittel ein wenig oberhalb der Vergleiche zwischen Übersetzungen und Originalen. Auffallend ist der sehr flache, dabei jedoch sehr lange rechte Schwanz der Verteilung. Die Textpaare, die hier zu finden sind, enthalten entweder einen der sehr kurzen Texte, oder sind thematisch untereinander eng verwandt.[44]

Tabelle 3.2 und Abbildung 3.11 geben einen Überblick über die Performanz des Algorithmus auf allen 11 Textvarianten. Die Werte liegen für alle Repräsentationen außer der `tag`-Repräsentation deutlich oberhalb der einfachen SVM-Klassifikatoren von Baroni und Bernardini (2006). Die Mittelwerte liegen zwar unterhalb der Werte der besten 10 kombinierten SVM-Klassifikatoren, aber die Streuung ist so hoch, dass sich keine signifikanten Unterschiede ergeben. Für die `tok`-Repräsentation in der originalen Tokenreihenfolge (`seq`) ergibt sich im Vergleich zum von Baroni und Bernardini (2006) zitierten Maximalwert ($F = 0.862$) ein p-Wert von 0.20 ($t = 1.35$). Dies ist insofern ein positives

[43] Hier wäre die Tokenisierung nicht trivial gewesen.

[44] Auffällig sind vor allem zwei Texte. Einer handelt von der Situation in Exjugoslawien unter besonderer Berücksichtigung des Kosovo. Der andere behandelt die wirtschaftliche Situation Albaniens (Marta L. Spagniolo, Marco Baroni, priv. Komm.). Die hohe Ähnlichkeit wird möglicherweise durch die Tatsache verstärkt, dass beide Artikel teilweise dieselben Werke zitieren.

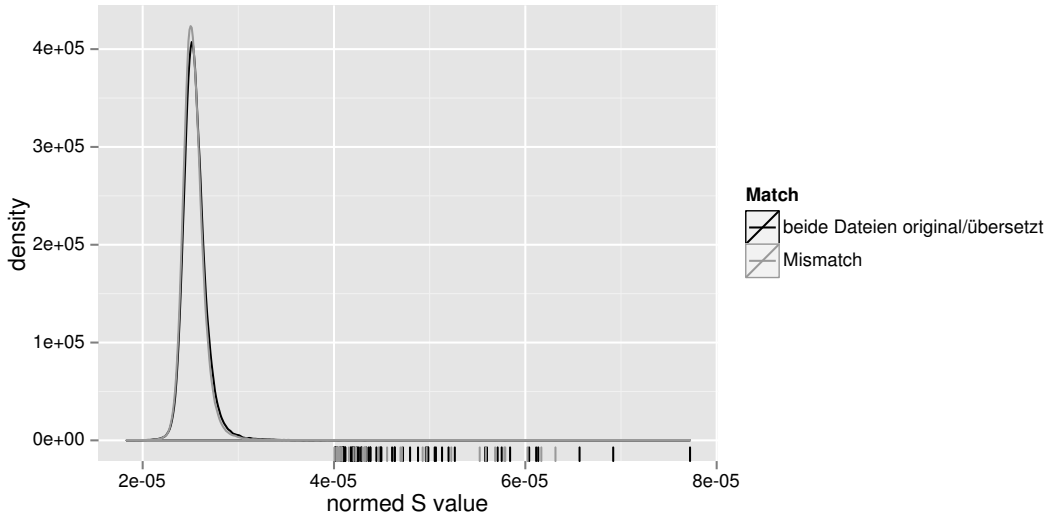


Abbildung 3.10: Die Verteilung der normierten S_{log}-Werte aller $\frac{813 \cdot 812}{2} = 330078$ Textvergleiche. Grundlage ist die *tok*-Darstellung. Die dunklere Kurve repräsentiert die Fälle, in denen beide Texte Übersetzungen oder beide Texte Originale waren (*Match*). Die hellere Kurve repräsentiert die übrigen Fälle. Bemerkenswert ist zum einen der extrem kleine Vorteil für die *Match*-Bedingung. Dieser winzige Unterschied ist ausreichend, mehr als 80% der Texte korrekt zu klassifizieren. Auch der extrem lange rechte Schwanz ist eine auffällige Eigenschaft der Verteilung. Die senkrechten Balken zeigen die S-Werte $> 4 \cdot 10^{-5}$ an. Diese Datenpunkte betreffen vor allem einerseits sehr kurze Dateien und andererseits Texte mit einer überproportionalen thematischen Ähnlichkeit.

	seq	ran
`txt`	0.842 ± 0.061	
`tok`	0.84 ± 0.069	0.817 ± 0.077
`mix`	0.801 ± 0.07	0.741 ± 0.072
`fun`	0.826 ± 0.067	0.814 ± 0.078
`lex`	0.811 ± 0.077	0.801 ± 0.07
`tag`	0.59 ± 0.074	0.529 ± 0.123

Tabelle 3.2: Mittelwerte der F-Werte unter S_{log} in den verschiedenen Textrepräsentationen. Der zitierte Fehler ist das Konfidenzintervall eines t-Tests über die 16 Durchläufe *Cross Validation*. Von der `tag`-Repräsentation abgesehen, wo beide Werte fast identisch sind, überstieg der *Recall* konsistent die *Precision*. Dies war in der Arbeit von Baroni und Bernardini (2006) genau umgekehrt.

Ergebnis, als weder ein modernes maschinelles Lernverfahren wie SVM verwendet wurde,

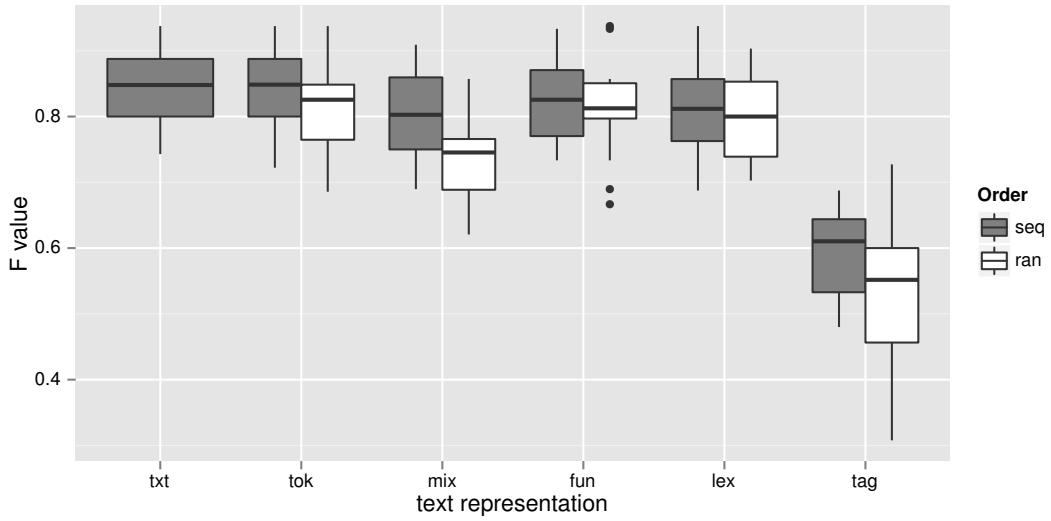


Abbildung 3.11: Visualisierung der Daten in Tabelle 3.2. Der Vergleich der Performanz (F-Wert) in den verschiedenen Repräsentationen der Texte. Die dunklen Boxen repräsentieren die originale Reihenfolge der Token, die weißen Boxen beziehen sich auf die randomisierten Texte.

noch linguistisches Wissen in Form von POS-Tagging oder Lemmatisierung eingegangen ist. Darüber hinaus geht nur eine einzige Repräsentation der Daten in die Analyse ein.

Auffällig ist, dass in jedem Fall die randomisierte Tokenreihenfolge schlechter abschneidet als die Originalreihenfolge. Für `tok`, `mix` und `tag` ist der Unterschied signifikant.[45] Dies steht in gewissem Gegensatz zu den Ergebnissen von Baroni und Bernardini (2006), die beobachten, dass es sich vor allem für die Repräsentationen mit lexikalischem Gehalt nicht auszahlt, n-Gramme mit $n > 2$ zu verwenden. In der vorliegenden Untersuchung hingegen ist die Kettenlänge unbeschränkt. Auffällig ist, dass die Randomisierung der Texte in der `mix`-Repräsentation einen stärkeren Effekt zu haben scheint als in den übrigen Repräsentationen. Auf der derzeitigen Datengrundlage kann nicht entschieden werden, ob dieses Phänomen stabil ist.

Das Abschneiden der `tag`-Repräsentation liegt sehr deutlich unter den übrigen Textversionen: In randomisierter Reihenfolge (`ran`) besteht kein signifikanter Abstand mehr von der Baseline 1/2. Das vergleichbarste Modell, das Baroni und Bernardini (2006) untersuchen, ist das *POS*-Unigram-Modell. Sie messen hier einen F-Wert nahe Null, da so gut wie keines der Testdokumente als Übersetzung klassifiziert wird. Tendenziell bestätigen sich hier unsere Ergebnisse somit gegenseitig. Für die Originalreihenfolge (`seq`) dagegen ergibt sich auch für `tag` ein signifikanter Unterschied zur Baseline. Das ist durchaus bemerkenswert, handelt es sich doch jeweils um eine Abfolge von POS-Tags lediglich der inhaltstragenden Wörter.

Weiterhin kann in Tabelle 3.2 abgelesen werden, dass alle Repräsentationen außer `tag`

[45] gepaarte t-Tests über alle 16 *Cross Validation Runs.*

vor allem in der Originalreihenfolge (`seq`) so gut wie identisch abschneiden. Dies kontrastiert mit den Ergebnissen von Baroni und Bernardini (2006), dargestellt in Abbildung 3.9. Ihre Ergebnisse streuen zwischen den Repräsentationen um jeweils mindestens 0.2.

Man kann diese gleichmäßigen Ergebnisse so interpretieren, dass es im Korpus eine Performanzgrenze um $F = 0.85$ gibt. Außer `tag` enthalten alle Repräsentationen ausreichend Information, um diese Grenze im Wesentlichen zu erreichen. Die Tatsache, dass Baroni und Bernardini (2006) mit einem maximalen F von 0.862 in einem sehr ähnlichen Bereich liegen, stützt diese These. Bemerkenswert ist, dass diese gleichmäßige Performanz einerseits auf sehr unterschiedlichen Verfahren beruht (SVM und S-basierte Klassifikation) und auch die von mir getesteten Repräsentationen teilweise komplementäre Information enthalten. So besteht `fun` ausschließlich aus den Oberflächenformen der Funktionswörter, `lex` dagegen aus den Oberflächenformen der inhaltstragenden Wörter. Dennoch lassen beide eine fast identische Klassifzierung in Übersetzungen und Originale zu. Diese Beobachtungen suggerieren die Modifizierung einer Aussage von Baroni und Bernardini (2006, 26):

> [...] while lexical cues help, they are by no means necessary, and translated text can be identified purely on the basis of function word distributions and shallow syntactic patterns.

Hier stellt sich die Situation eher so dar, dass lexikalische Information zwar tatsächlich nicht notwendig ist, aber für sich genommen ausreichend um die Texte beinahe so genau zu klassifzieren wie die volle Folge aller Oberflächenformen (`txt`).

Einen weiteren Aspekt dieses Phänomens zeigt das linke Teilbild von Abbildung 3.12. Hier sind die F-Werte der einzelnen *Cross Validation Runs* dargestellt. Das linke Teilbild zeigt das logarithmische Maß S_{log}. Betrachten wir zuerst das Verhältnis der Performanz in den Repräsentationen `tag` und `lex`. Dass `tag` wesentlich schlechter abschneidet ist bereits bekannt. Es ist aber auch zu erkennen, dass in den *Cross Validation Runs*, in denen `lex` besonders gut abschnitt, die Performanz in `tag` überdurchschnittlich tief liegt und umgekehrt. Der Zusammenhang ist mit den vorhandenen 16 Datenpunkten allerdings nicht signifikant.

Ähnliche Auffälligkeiten deuten sich auch im Verhältnis der Repräsentationen `tok`, `mix` und `fun` an. In 7 der 16 Fälle ergibt sich ein Muster, dass `tok` und `fun` relativ ähnlich abschneiden, `mix` aber in der einen oder anderen Richtung stärker abweicht. Insgesamt zeichnet sich das Bild ab, dass die verschiedenen *Cross Validation Runs* in verschiedenen *Repräsentationen* unterschiedlich abschneiden und dass es systematische Korrelationen und Gegenläufigkeiten gibt. Von Durchlauf zu Durchlauf unterscheiden sich die Textfiles. Es scheint somit so als würden in den unterschiedlichen Texten die Hinweise auf ihren Charakter als Übersetzung oder Original auf unterschiedlichen Annotationsebenen liegen.

In der bisherigen Forschung zum Thema wird von einigen Autoren (Baker, 1993; Ilisei et al., 2010; Laviosa, 1998, 2002) ein *Simplification Universal* vorgeschlagen und empirisch untersucht, von Ilisei et al. (2010) beschrieben als „the tendency of translators to produce simpler and easier-to-follow texts (Baker, 1993)“. Baroni und Bernardini

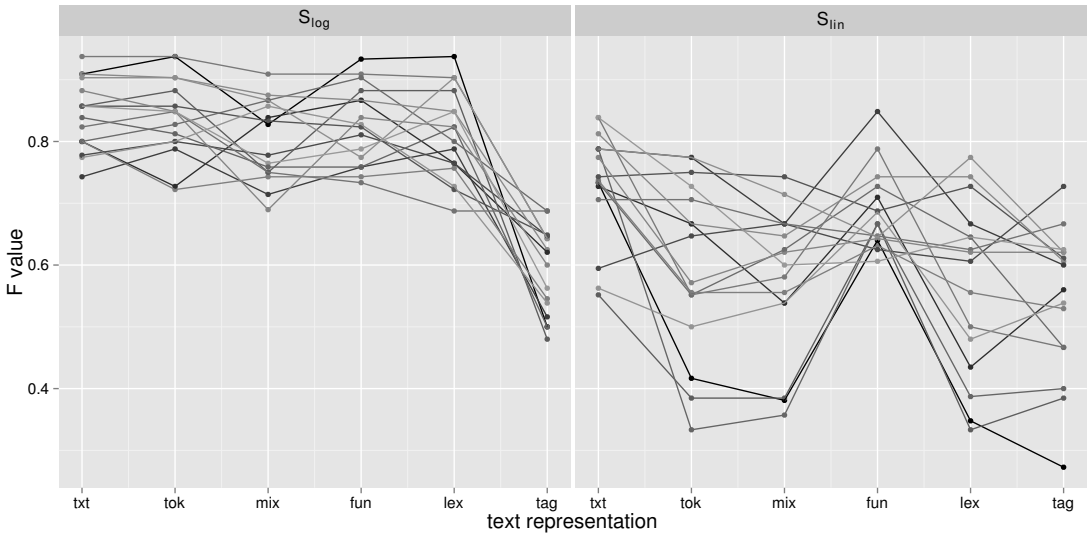


Abbildung 3.12: Die Linien identifizieren jeweils einen der 16 Cross-Validation-Durchläufe in den verschiedenen Repräsentationen. Datengrundlage sind die Texte in ihrer Originalreihenfolge. Das linke Teilbild zeigt das logarithmische Maß S_{log}, rechts ist das lineare S_{lin} dargestellt.

(2006) identifizieren „the distribution of function words and morphosyntactic categories in general, and personal pronouns and adverbs in particular“ als die Eigenschaften eines Textes, die eine Klassifizierung in Übersetzungen und Originale erlauben. Diesen unterschiedlichen Erklärungsansätzen ist gemeinsam, dass sie von *Translationese* als einem homogenen Phänomen ausgehen.

Würden weitere empirische Fakten der von meinen Ergebnissen suggerierten Hypothese mehr Gewicht verleihen, dass sich verschiedene Übersetzungen in verschiedenen Punkten von Originalen unterscheiden, würde dies ein neues Licht auf die *Translationese*-Debatte werfen.

Diese Untersuchung fügt auch der auf Seite 135 dargestellten Diskussion einen neuen Aspekt hinzu. Dort ist die in der Literatur populäre Ansicht dargestellt, dass es schädlich sein könnte, den unveränderten Text einschließlich der Inhaltswörter als Datengrundlage zu nehmen, da dadurch schwer zu kontrollierende Wechselwirkungen mit dem Inhalt des Textes auftreten können. Zumindest im vorliegenden Korpus ist es nun so, dass das Vorhandensein dieser Information auf der einen Seite zu keiner Herabsetzung der Performanz führt und dabei selbst ausreichend ist, die vorliegende stilometrische Fragestellung zu beantworten. Allerdings ist das vorliegende Korpus auf größtmögliche *Topic*-Homogenität hin zusammengestellt. Daher sind auch die Folgen größerer Heterogenität des *Topics* daran nicht abschätzbar.

Insgesamt bleibt der Eindruck bestehen, dass in der bisherigen *Translationese*-Forschung – wie vielleicht in der Stilometrie allgemein – die Nützlichkeit lexikalischer Information im Vergleich zu Funktionswörtern und POS-tag-Folgen unterschätzt wird. Das

rechte Teilbild von Abbildung 3.12 legt hierfür eine Begründung nahe. Hier ist das Verhalten des linearen Maßes S_{lin} dargestellt. Nun schneiden die Repräsentationen stark unterschiedlich ab. Am zuverlässigsten erlaubt die `fun`-Repräsentation eine Klassifzierung der Texte. Dies wiederum ist im Geiste der bisherigen stilometrischen Forschung, die ja zu einem großen Teil auf der Untersuchung der Funktionswörter fußt (3.2, Seite 135).

Im Unterschied zu S_{log} gehen in S_{lin} alle Frequenzen linear ein. Dies führt zu einer höheren Bewertung der kurzen und damit häufigen Ketten, da die längeren Elemente um so viele Größenordnungen seltener sind, dass sie keine Rolle mehr spielen können. Da gerade die inhaltstragenden Wörtern in natürlichsprachigen Texten so besonders ungleichmäßig verteilt sind, geht ein großer Teil der Information verloren, wenn alle seltenen Wörter unbeachtet bleiben. Dieser Teil ist wertvoll, wie sich am gleichmäßig guten Abschneiden des logarithmischen Maßes zeigt. Es scheint nachvollziehbar, dass dieses Problem innerhalb der geschlossenen Klasse der Funktionswörter nicht so stark zum Tragen kommt.

Ein weiterer Punkt fällt am rechten Teilbild von Abbildung 3.12 auf. Mit zwei Ausnahmen schneidet die `txt`-Repräsentation besser ab als `tok`, teilweise liegen die F-Werte um volle 0.4 auseinander. Dies steht in scharfem Gegensatz zum logarithmischen S_{log} im linken Teilbild, wo die Varianz beider Repräsentationen gering ist. `tok` und `txt` unterscheiden sich nur darin, dass in `tok` die Eigennamen durch neutrale Platzhalter ersetzt wurden, da sie möglicherweise triviale Hinweise darauf geben könnten, was eine Übersetzung und was ein Original ist. Die diskutierten Ergebnisse weisen darauf hin, dass dies tatsächlich der Fall ist, falls man mit einem Ähnlichkeitsmaß arbeitet, dass den häufigen Elementen ein ungebührlich hohes Gewicht gibt.

Eine weitere interessante Frage ist die Abhängigkeit der Klassifikationsqualität von der Länge des Trainingskorpus. Wie viel Beispielmaterial braucht der Algorithmus, um zwischen Übersetzung und Original unterscheiden zu können? Die Ergebnisse sind in Abbildung 3.13 dargestellt. Eine erste, die bisherigen Ergebnise stützende Beobachtung ist, dass die in Ungleichung 3.5 dargestellte Reihenfolge deutlich reproduziert wird. Die Endpunkte der Kurven enden in drei Gruppen: Die doppellogarithmischen Maße S_{log} und S_{mlog} schneiden am besten ab, mit einem hauchdünnen Vorteil für S_{log}. Dann folgen die drei halblogarithmischen Maße S_{rlog}, S_{llog} und S_{shlog}. Abgeschlagen schließlich das lineare S_l.

Ab vielleicht 10000 Zeichen, etwa einigen Seiten Text, steigt die Performanz stark an. Noch am Endpunkt bei 10^5 Zeichen liegt die *Accuracy* ein gutes Stück unter dem im Gesamtkorpus (einige Millionen Zeichen) erreichten Wert. Die Güte der Klassifikation hängt also im untersuchten Bereich stark von der Menge des zur Verfügung stehenden Textes ab.

Überraschend ist, dass die Kurve für S_{log} (und S_{llog}) schon ganz zu Beginn, bei 20 Zeichen, über der Baseline liegt: Der Durchschnitt von 0.511 liegt signifikant über 50% ($t = 2.60$, $p = 0.01$).

Interessant ist der Verlauf der halblogarithmischen Maße in Verhältnis zum linearen S_l auf der einen Seite und den doppellogarithmischen auf der anderen Seite: Zu Beginn verläuft die Kurve von S_{llog} parallel zu S_{log} um dann gegen Ende nach unten abzufallen. Umgekehrt verhält es sich mit S_{rlog} und S_{shlog}, die unten bei S_l starten um dann gleichauf

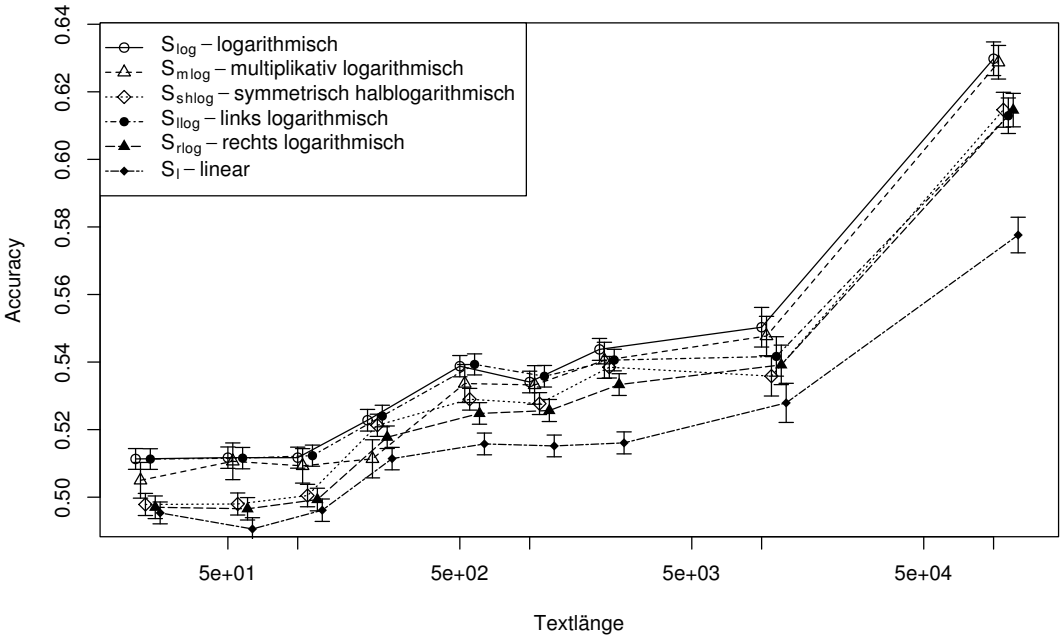


Abbildung 3.13: Wie hängt die Performanz der verschiedenen Maße von der Trainingstextlänge ab? Datengrundlage ist die `txt`-Repräsentation.

mit S_{llog} zu enden.

T_1 ist der Trainingstext, während T_2 den Testtext bezeichnet. Daher ist S_{llog} logarithmisch in den Frequenzen des Testtextes, der zu Beginn viel länger ist. S_{rlog} dagegen ist linear in Bezug auf den Testtext und logarithmisch in Bezug auf den Trainingstext, der zu Beginn sehr kurz ist. Daraus kann man den Schluss ziehen, dass das Logarithmieren der Frequenzen nur dann einen Vorteil bringt, wenn der entsprechende Text lang genug ist. Aus der Abbildung kann man schätzen, dass es dazu einige Hundert Zeichen braucht.

Der Verlauf der Kurven relativ zueinander scheint recht regelmäßig, während der Verlauf der Kurven insgesamt ein bisschen chaotisch aussieht. Betrachtet man den Verlauf für die drei durchgeführten Randomisierungen des Korpus getrennt (hier nicht gezeigt), so zeigt sich jeweils ein unterschiedlicher Verlauf der Kurven insgesamt, während die im vorhergehenden Absatz beschriebenen relativen Verhältnisse sich gleich bleiben. Dies könnte sich daraus erklären lassen, dass manche Trainingsfiles besonders gute Beispiele für ihre Textart sind, während andere den Algorithmus stark in die Irre führen. Sobald diese besonderen Dateien in das Trainingskorpus aufgenommen werden, erscheinen Sprünge in der Performanz. Bei jeder Randomisierung erscheinen diese Sprünge bei unterschiedlichen Trainingskorpuslängen. Ein sehr ähnliches irreguläres Verhalten haben Clement und Sharp (2003, Fig. 8) festgestellt. Auch dort schwankt die Performanz stark mit der Randomisierung.

3.6.2 Klassifikation von Lernertexten nach der Muttersprache des Autors

Dieser Abschnitt setzt sich damit auseinander, ob und wie gut mit Hilfe von S die Muttersprache des Autors eines Textes vorhergesagt werden kann.

Mit dieser Untersuchung ist nach Autorenbestimmung in 3.5 und *Translationese* in 3.6.1 eine weitere stilometrische Aufgabenstellung Thema. Stilometrische Performanzwerte schwanken gemeinhin stark von Aufgabenstellung zu Aufgabenstellung und von Datensatz zu Datensatz. Je mehr unabhängige Fragestellungen an unterschiedlichen Datensätzen mit einer neuen Methode untersucht werden, desto größer kann das Vertrauen in die gewonnenen Ergebnisse sein.

Eine weitere Motivation für diese Untersuchung war wie bereits in Abschnitt 3.6.1 die Tatsache, dass die Ergebnisse auch hier mit etablierten und experimentell sauber durchgeführten Arbeiten verglichen werden können, die dieselbe Frage am selben Datensatz bearbeiten.

Im Laufe der Untersuchung zu Tage tretende Fehler im Korpus und systematische Eigenheiten der Daten erzwangen auch in diesem Zusammenhang eine Auseinandersetzung mit dem Problem der Korpusreinheit. Es zeigt sich, dass gerade die hier untersuchten Daten und Verfahren geeignet sind, übermäßige Wiederholungen in einem größeren Korpus zu finden und dieses zu säubern. Dadurch wird die Qualität der Klassifikation wesentlich erhöht. Ähnliche Beobachtungen werden auch in Abschnitt 65 im Zusammenhang mit dem dort verwendeten türkischen Korpus (Say et al., 2002) berichtet.

Auch an diesem Korpus lässt sich die nun bereits etablierte Performanzreihenfolge der verschiedenen S-Maße weiter untermauern.

Die Literatur zu diesem Bereich der Stilometrie ist so überschaubar, dass man sie hier vollständig erwähnen kann. Den ersten Versuch, die Muttersprache eines Autors aus englischen Texten zu ermitteln, haben Koppel et al. (2005, 2006) unternommen.

Koppel et al. haben gezeigt, dass diese Aufgabe mit verblüffend hoher Genauigkeit gelöst werden kann, wenn man intensiv annotierten Text und ein effektives maschinelles Lernverfahren verwendet (*Support Vector Machines* (SVM), vergleiche 3.2). Sie beziehen nicht nur die Verteilungen von n-Grammen, POS-Tags und Funktionswörter mit ein, sondern verbessern ihre Methode durch die Berücksichtigung verschiedener Arten von Fehlern.

Koppel et al. verwenden für ihre Untersuchungen das *International Corpus of Learner English* (ICLE), zusammengestellt von Granger (2003), ein Korpus mit Texten von fortgeschrittenen Englischlernern. Sie beschränken sich für ihre Untersuchungen auf eine Untermenge von 5 Sprachen: Bulgarisch (BG), Tschechisch (CZ), Spanisch (SP), Russisch (RU) und Französisch (FR). An dieser Stelle ist zu erwähnen, dass dies gegenüber der im vorigen Kapitel bearbeiteten Fragestellung eine Erweiterung darstellt, da nun 5 und nicht mehr nur 2 Kategorien existieren.

Mit ihrem Verfahren erreichen sie eine *Accuracy* von 80.2%. Dieser Wert bezieht den vollen Umfang der untersuchten Daten mit ein. Verwenden sie nur Zeichen-n-Gramme, erreichen sie nur knapp unter (ohne Fehler) oder knapp über (mit Fehlern) 70%. Welche n-Gramme sie genau verwendet haben, wird nicht erwähnt, nur dass es 200 waren. Es zeigt sich, dass die hier vorgestellte Methode unter Ausnutzung der vollständigen Fre-

quenzen aller n-Gramme eine Genauigkeit erreicht, die fast identisch ist mit der von Koppel und Kollegen maximal erreichten.

Tsur und Rappoport (2007) beziehen sich auf diese ursprüngliche Arbeit und weisen nach, dass bereits die 84 häufigsten Bigramm-Frequenzen im Verein mit *support vector machines* ausreichen, um Genauigkeiten deutlich über der Baseline zu bekommen (66%).

Seit diesen Anfängen gab es meines Wissens nur noch zwei Arbeiten zum Thema.

Der Schwerpunkt von Estival et al. (2008) liegt auf dem *author profiling*. Die Bestimmung der Muttersprache ist in dieser Arbeit nur ein Nebenaspekt. Verschiedene maschinelle Lernverfahren werden anhand von *Feature Vectors* verglichen. Da sie ein eigenes Korpus verwenden und andere Sprachen untersuchen, sind ihre Ergebnisse mit der aktuellen Untersuchung nicht vergleichbar.

JojoWong und Dras (2009) untersuchen den Einfluss syntaktischer Fehler auf das Klassifikationsergebnis. Diese Fehlerklasse wurde von Koppel et al. (2005, 2006) ausgespart. Sie beobachten keine wirkliche Verbesserung. Sie verwenden zwar das *ICLE*-Korpus, aber eine andere Version und eine andere Sprachauswahl, so dass auch ihre Ergebnisse sich nicht direkt auf die ursprünglichen Arbeiten beziehen lassen.

Für die hier berichtete Untersuchung wurden 12 *cross validation*-Durchläufe durchgeführt. Es wurden alle 12 Sprachen einbezogen, nicht nur die oben erwähnte Untermenge. Zusätzlich sind das: Flämisch (Belgien, DB), Niederländisch (DN), Finnisch (FI), Deutsch (GE), Italienisch (IT), Polnisch (PO) und Schwedisch (SW). In jedem Durchlauf wurden 10 Dateien aus jedem der 12 Unterkorpora als Testfiles benutzt, die übrigen bildeten die Trainingskorpora. Insgesamt wurden also $12 \cdot 12 \cdot 10 = 1440$ Dateien klassifiziert.

Für den vollen Datensatz ergibt sich eine *Accuracy* von 0.647 ± 0.008. Eingeschränkt auf den Teil des Datensatzes, der von (Tsur und Rappoport, 2007; Koppel et al., 2005, 2006) verwendet wurde (BG, CZ, SP, RU und FR), ergeben sich 0.74 ± 0.02.

Wenn man bedenkt, dass die Daten von Koppel et al. reich annotiert waren und dass diese ein etabliertes maschinelles Lernverfahren verwendet haben, ist dies bereits ein ermutigendes Ergebnis.

Ich komme nun zum wichtigen Thema der Korpusreinheit. Unter diesem Begriff fasse ich verschiedene mögliche Fehlerquellen zusammen. Eine davon hatten wir bereits im dem Metu-Korpus (Say et al., 2002) in Abschnitt 65 kennengelernt. Dort treten vollständige Wiederholungen von Textfragmenten auf. So etwas ist vor allem bei webbasierten Korpora schwer zu vermeiden.

Aber auch bei ICLE, einem Essaykorpus, ergeben sich bei genauerer Inspektion der Daten verdächtige Muster. Siehe dazu Abbildung 3.14. Für diese Graphik wurde jede Datei T_i mit allen anderen Dateien T_j $(i \neq j)$ aus demselben Unterkorpus mit Hilfe des Ähnlichkeitsmaßes $S_{log}(T_i, T_j)$ verglichen. Für ein festes i wurde der Mittelwert über alle j gebildet und jeder S-Wert durch den entsprechenden Mittelwert geteilt. In Abbildung 3.14 sind die so reskalierten S-Werte ihrer Größe nach geordnet. Im Bereich kleinerer Abweichungen der S-Werte vom Mittelwert (rechts), fällt die Kurve sehr regelmäßig ab. Auf der linken Seite erfolgt etwa bei Rangplatz 35 erst ein leichter Knick nach oben, gefolgt von plötzlich noch viel größeren Werten.

Für die zu diesen links liegenden S-Werten gehörenden Dateipaare ergeben sich Un-

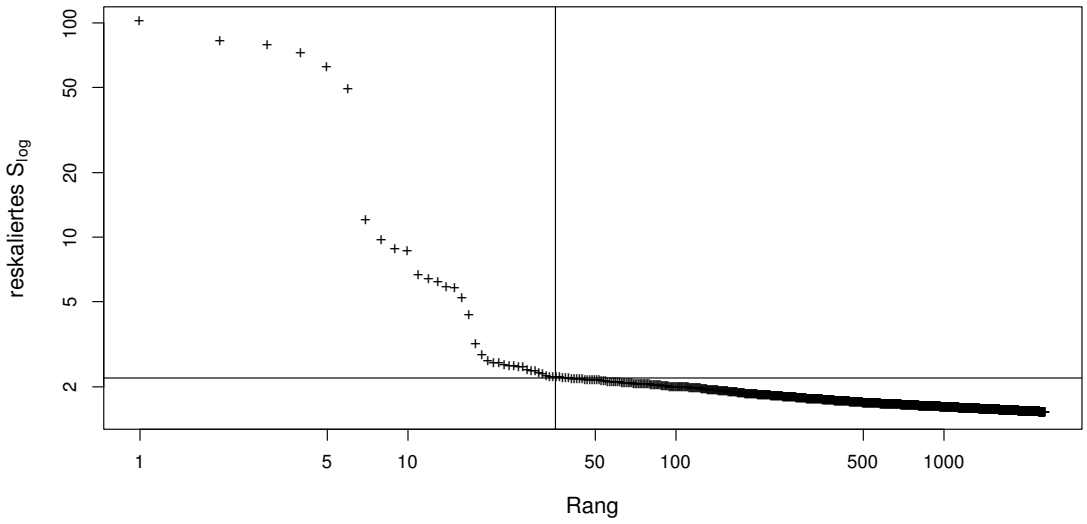


Abbildung 3.14: Verteilung der reskalierten S-Werte. Die Darstellung ist doppellogarithmisch. Die gekreuzten Linien bezeichnen den Übergang zwischen einer regelmäßigen Verteilung (rechts) und abnormal hohen Werten (links). Weitere Details im Text.

gereimtheiten: Es gibt mindestens 12 volle Dubletten. So sind die Dateien `SWUG2028` und `SWUG2040` identisch. Die Metadaten zu diesen Dateien sind ebenfalls beinahe identisch, nur fehlt in einem Fall das Alter der Autorin.

Daneben gibt es auch starke Kopien. So endet Datei `DNNI4012` mit 1090 Zeichen aus Datei `DNNI4008`. Hier sind die Metadaten völlig unterschiedlich.[46]

Ein von diesen groben Fehlern unabhängiges Problem ist das folgende: Oft merkt man den Texten erheblich den Einfluss der Themenstellung an, dh. ganze Sätze oder Abschnitte wiederholen sich von Text zu Text. Es findet sich auch häufig der Text der Aufgabenstellung unverändert im Text. Das ist zwar normal und zu erwarten. Nichtsdestotrotz stellt es ein Problem dar, weil es zu künstlichen Ähnlichkeiten führen kann.

Einen Eindruck des Problems vermittelt Abbildung 3.15. Es zeigt eine Übersicht über die Aufsatzthemen, verteilt nach der Zahl der Länder, in denen jedes Thema vergeben wurde.[47] Zwar wurde die Mehrzahl der Themen genau so nur ein einziges Mal vergeben. Für das übrige Viertel der Fälle häuft sich aber die Situation, dass dieselbe Aufgabenstellung vor allem innerhalb einer Muttersprachengruppe bearbeitet wurde. Dies sorgt für eine unnatürlich hohe Ähnlichkeit der Texte.

[46] Ich habe die Texte von der CD (Granger, 2003) herunterkopiert, statt das mitgegebene Analyseprogramm zu verwenden. Ein kleiner Teil dieser Probleme taucht nicht auf, wenn man die installierte CD wie vorgesehen durchsucht. Die Datei `FRUC4034`, die ein Torso von `FRUC3034` ist, taucht in den Metadaten nicht auf. Die meisten Probleme finden sich aber auch, wenn man auf die Daten von der Suchmaske aus zugreift.

[47] Gemäß den Metadaten.

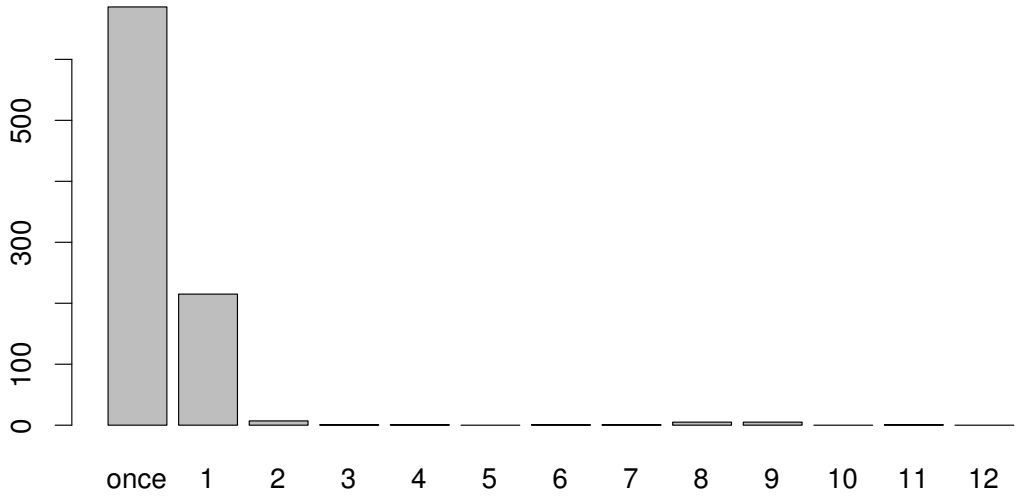


Abbildung 3.15: Verteilung der Aufsatzthemen in ICLE. Die meisten Themenstellungen erscheinen nur einmal (74.3%). Viele der übrigen werden zwar mehr als einmal vergeben, aber nur innerhalb eines einzigen Landes (23.3%). Nur die restlichen 2.4% der Titel wurden in mehr als einem Land vergeben. Kodierungsfehler, aufgrund derer derselbe Titel nicht wiedererkannt wird sind möglich, einer manuell überprüften Stichprobe nach aber nicht häufig.

Diese Probleme wurde meines Wissens von den zitierten Autoren weder bemerkt noch korrigiert und könnten durchaus dazu in der Lage sein, die Performanz dieser Ansätze zu verbessern. Dies gilt wohl besonders für die von Tsur und Rappoport (2007) verwendeten Bigramme, welche für lexikalische Einflüsse anfällig sind.

Für die S-basierte Stilometrie sind übermäßige Wiederholungen eher schädlich, da die Normalisierung der S-Matrizen von einer Verteilung ausgeht, die keine stärkeren Ausreißer hat. Die Performanz der Methode steigt mit der Reinheit des Korpus an.

Die Dubletten lassen sich manuell identifizieren und entfernen. Das restliche Korpus wird von allen übermäßigen Wiederholungen gereinigt: Alle Stellen von über 30 Zeichen Länge, die sich im gleichen Subkorpus wiederholen, werden entfernt.[48]

Die Reinigung verbessert die Ergebnisse erheblich: Die Accuracy steigt auf 0.705 ± 0.017 für das volle Korpus und 0.79 ± 0.01 für den 5-Sprachen-Teil. Die zitierten Fehler sind der *Standard Error of the Mean* (SEM).

Das Ergebnis für die 5-Sprachen-Untermenge ist nun fast ununterscheidbar vom Wert, den Koppel et al. (2005, 2006) erhalten haben. Hier werden aber ausschließlich Substringhäufigkeiten herangezogen und die Klassifikation erfolgt parameterfrei durch den

[48] Ein sehr ähnliches Problem ergibt sich mit einem anderen, in Abschnitt 3.6.4 untersuchten Korpus. Dort lohnt es sich, ein wesentlich ausgeklügelteres Verfahren einzusetzen um wörtliche Kopien aus der Themenstellung zu erkennen und zu beseitigen.

Vergleich von S-Werten. Koppel et al. dagegen ziehen wesentlich mehr und vielseitigere Information zu Rate und klassifizieren mittels *Support Vector Machines* (SVM). Es ist vorstellbar, dass die S-basierte Klassifikation mit einer sorgfältigeren Reinigung des Korpus noch ein wenig bessere Klassifikationsergebnisse liefert. Substanzielle Verbesserungen erwarte ich nicht.

Es ist in diesem Zusammenhang eine interessante Tatsache, dass ich nur Wiederholungen *innerhalb* einer Muttersprache entferne, nicht *zwischen* den Sprachen. Damit werden unnatürlich starke Ähnlichkeiten *herausgenommen*, man könnte daher vermuten, dass die Ergebnisse eher *schlechter* werden. Tatsächlich werden sie *besser*. Dies zeigt einerseits die Wichtigkeit und Sensibilität der Normalisierung von S, lässt aber auch Raum für die Möglichkeit, dass die Methoden von Tsur und Rappoport (2007); Koppel et al. (2006) auf dem gereinigten Korpus schlechter arbeiten würden, da sie von den Wiederholungen profitieren könnten. Es gibt dort keinen Schritt, der dem der Normalisierung entsprechen würde.

Das vorgestellte Verfahren übersieht diese Unstimmigkeiten nicht nur nicht, sondern hilft zugleich sie zu neutralisieren. Hier zeichnen sich hilfreiche Anwendungen ab.

Auch nach der beschriebenen Reinigung zeigen die Daten eine auffällige Eigenschaft: Es gibt eine Häufung der Missklassifikationen von Autoren verschiedener Muttersprachen in die bulgarische Sprechergruppe (Siehe Abbildung 3.16). In den von Koppel et al. zitierten Ergebnissen findet sich ein ähnliches Muster, auch dort bilden die als Bulgarisch klassifizierten Dateien die größte Gruppe und auch für die übrigen Sprachen ergeben sich ähnliche Verhältnisse wie hier, aber die Effekte sind nicht so ausgeprägt.

Dies legt den Schluss nahe, dass das Ausgangsproblem in den Daten liegt, aber nur die hier vorgestellte Methode in der Lage ist, es in den Ergebnissen deutlich erkennbar zu reflektieren.

Die Subkorpora zu den verschiedenen Muttersprachen sind unterschiedlich groß. Man könnte vermuten, dass das sehr unterschiedliche Maß an Fehlklassifikationen und vor allem der Ausreißer BG damit zusammenhängen. Eine Auftragung der Zahl der Missklassifikationen über der Korpusgröße (Abbildung 3.17) ergibt allerdings keinen Zusammenhang. Woran die auffälligen Ergebnisse für Bulgarisch liegen, bleibt unklar. Ich habe keine Eigenschaften gefunden, die das BG-Subkorpus auszeichnen. Ein Ansatzpunkt für weitere Überlegungen könnte folgendes Zitat darstellen: „For example, the Bulgarian authors were on average considerably less prone to errors than the Spanisch authors.“ (Koppel et al., 2005, 4)

Bemerkenswert ist in diesem Zusammenhang die erhebliche Verbesserung der Ergebnisse, wenn BG außen gelassen wird. Es ergibt sich eine Accuracy von 0.77 ± 0.02 für das Gesamtkorpus und 0.855 ± 0.007 für das 5-Sprachen-Subkorpus. Dies ist wesentlich oberhalb der Zahlen der Vergleichsarbeiten. Da diese in ihren Daten das BG-Problem nicht oder kaum sehen, ist es eine plausible Hypothese, dass das hiesige Verfahren für unproblematischere Daten besser funktionieren würde.

Die bisherigen Ergebnisse von mir und den Vergleichsarbeiten sind in Tabelle 3.3 zusammengefasst.

Es ist wieder interessant, sich noch einmal anzuschauen, wie die verschiedenen Varianten von S im Vergleich bei dieser Klassifikationsaufgabe abschneiden. Tabelle 3.4 zeigt

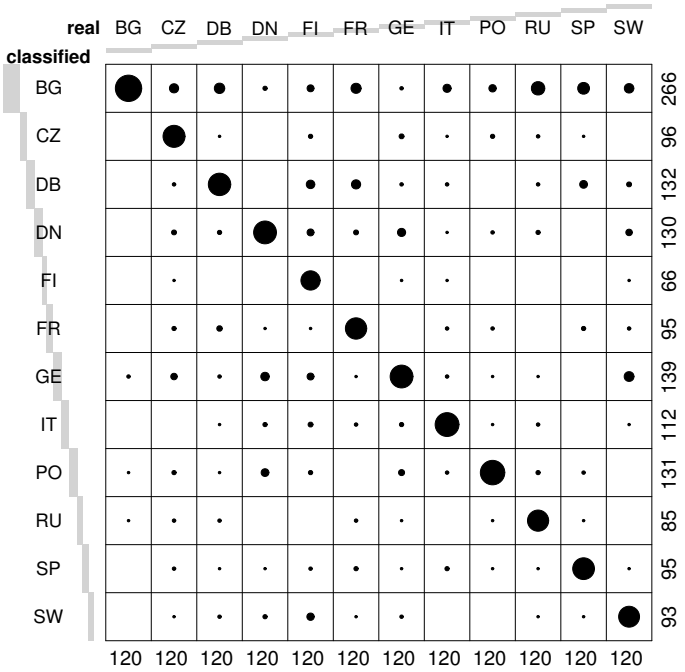


Abbildung 3.16: Balloonplot für die Klassifikationsergebnisse der ICLE-Texte nach der Muttersprache der Autoren. Die Spalten zeigen die tatsächliche Muttersprache an, die Reihen die Klassifikationsresultate. Die grauen Balken visualisieren die jeweiligen Reihen- und Spaltensummen, die unten und rechts noch einmal explizit angegeben sind. Der breite graue Balken bei Bulgarisch (BG) bedeutet, dass mehr als doppelt so viele Dateien als zu Bulgarisch (BG) gehörig klassifiziert werden, als wirklich im Korpus vorhanden sind. Andere Eigenschaften der Verteilung sind unauffällig und erwartbar. Zum Beispiel die häufige Verwechslung von Deutsch (DE) und Schwedisch (SW) oder auch von DE und Niederländisch (DN). Die Ähnlichkeit von Flämisch (DB) und Französisch (FR) ist interessant. Ebenso das Paar Schwedisch (SW) und Finnisch (FI).

die Ergebnisse. Diese decken sich mit der Reihenfolge wie wir sie bereits aus 3.6.1 und 3.5 kennen.

Wieder ergeben sich drei Gruppen: Fast identisch schließen S_{log} und S_{rlog} ab, deutlich schlechter sind S_{llog} und S_{shlog}. Weit abgeschlagen ist wieder S_l.

Wieso verhalten sich die beiden halb-logarithmischen Maße S_{rlog} und S_{llog} so unterschiedlich? Die Erklärung geht aus den in Anschluss an Abbildung 3.13 gegebenen

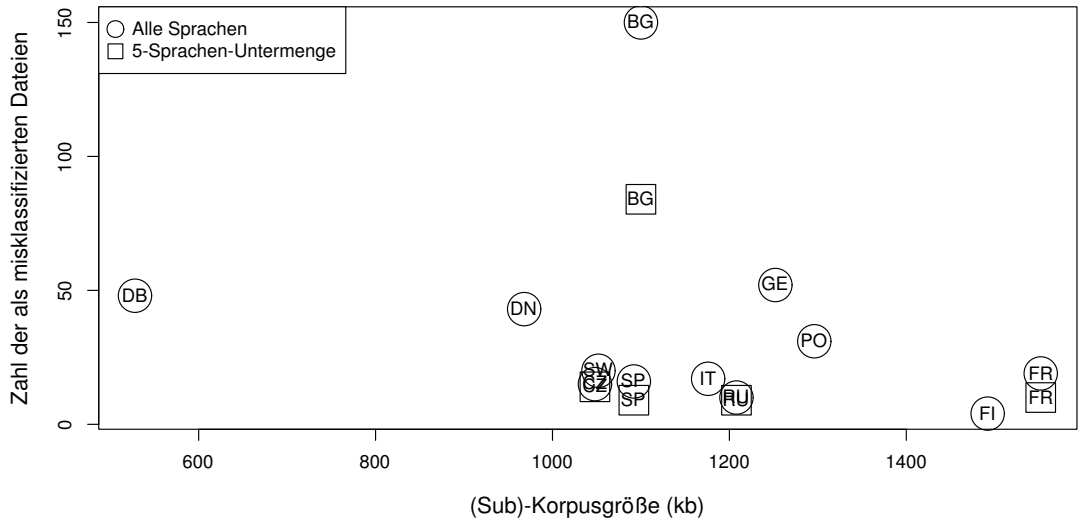


Abbildung 3.17: Abhängigkeit der Zahl der **fälschlich** in eine bestimmte Sprache einsortierten Dateien von der Korpusgröße. Man erkennt eine leicht abnehmende Tendenz mit der Korpusgröße. Bulgarisch verhält sich abweichend.

	Accuracy	
Beschreibung	12 Sprachen	Untermenge
ungefiltert	0.647 ± 0.008	0.74 ± 0.02
gefiltert	0.705 ± 0.017	0.79 ± 0.01
gefiltert ohne BG	0.77 ± 0.02	0.855 ± 0.007
Koppel et al. (2005)	–	0.802
Tsur und Rappoport (2007) | Buchst.-Bigr.		0.66
Tsur und Rappoport (2007) | Buchst.-Trigr.		0.5967
Tsur und Rappoport (2007) | Funktionswörter		0.667

Tabelle 3.3: Ergebnisübersicht. Angegebene Fehler jeweils SEM.

Erläuterungen hervor: S_{rlog} ist hier linear im viel kürzeren Testtext und logarithmisch im Trainingstext. Der Vorteil, der sich aus dem Logarithmieren ergibt, zeigt sich erst bei längeren Texten. Deswegen gibt es hier kaum einen Unterschied zwischen S_{rlog} und dem doppellogarithmischen S_{log}, während S_{llog} stark zurückfällt. Eine Kombination der beiden halblogarithmischen Maße gibt keine Vorteile, auch dies ist ein Ergebnis, das sich durchzieht.

In diesem Abschnitt wurde die vorgestellte stilometrische Textklassifikationsmethode an einer zweiten Fragestellung getestet: Bestimmung der Muttersprache der Autorinnen und Autoren englischer Lernertexte. Auch hier gab es wie im vorigen Abschnitt (*Translationese*, 3.6.1) Vergleichsmöglichkeiten mit Vorgängerarbeiten auf demselben Datensatz

Maß	Korrekt erkannt (von 1440)	Bemerkung
$S_{log}(trn, tst)$	1015	
$S_{rlog}(trn, tst)$	977	linear in tst
$S_{llog}(trn, tst)$	752	linear in trn
$S_{shlog}(trn, tst)$	746	
$S_{l}(trn, tst)$	604	
baseline	120	

Tabelle 3.4: Der Vergleich der S-Varianten. trn bezeichnet den Trainingstext, tst den Testtext. Der Unterschied zwischen den ersten beiden Maßen (S_{log} und S_{rlog}) und zwischen S_{llog} und S_{shlog} ist nicht signifikant (χ^2-Test). Die übrigen Unterschiede sind signifikant. Daten für S_{mlog} sind nicht vorhanden.

zur selben Fragestellung. Wieder konnte das Performanzmaximum der Vorgängermethoden reproduziert werden. Diese verwenden allerdings reich annotierte Daten und Maschinenlernverfahren, wogegen mein Ansatz lediglich *Zeichen-n*-Gramme heranzieht und die Klassifikation mittels eines reinen Zahlenvergleichs erfolgt. Auch die Performanzreihenfolge der S-Maß-Varianten konnte ein zweites Mal bestätigt werden. Es zeigte sich, dass das Korpus die Methode vor technische Probleme stellt, die einerseits aus Fehlern im Korpus selbst herrühren und andererseits aus der Systematik elizitierter Daten erwachsen. Die verwendeten vollständigen Substringfrequenzen geben uns aber Informationen an die Hand, um die problematischen Stellen aus dem Korpus herausfiltern zu können. Diese Filterung verbessert die Ergebnisse erheblich.

Mit den *Federalist Papers* wende ich mich nun einem Korpus zu, dass derartige Probleme nicht enthält. Dennoch zwingt die Struktur des Korpus dazu, die Methode selbst um einen weiteren Arbeitsschritt zu erweitern. Die erweiterte Methode bestätigt wiederum den Stand der Forschung.

3.6.3 Automatische Autorenbestimmung anhand der Federalist Papers

In diesem Kapitel wenden wir uns dem wohl am häufigsten behandelten Problem der Stilometrie überhaupt zu: Automatische Autorenbestimmung (*Authorship Attribution (AA)*) anhand der *Federalist Papers.*

Hierbei handelt es sich um 85 politische Aufsätze aus der Frühzeit der Vereinigten Staaten (1787-1788), deren Zweck es war, das Volk und vor allem die Legislative von New York von der noch zu ratifizierenden US-Verfassung zu überzeugen.

Die Artikel wurden unter einem Pseudonym veröffentlicht, hinter dem drei Autoren stehen: Alexander Hamilton, James Madison und John Jay. Bevor Hamilton bei einem Duell getötet wird, hinterlässt er eine Liste, die jedem Aufsatz einen Autor zuweist. Dieser in aller Hast entstandenen Liste hat Madison später (1818) – wohl mit Grund – widersprochen. In der Folge war für 12 der Artikel ungeklärt, ob sie von Madison oder Hamilton verfasst worden waren. Erst seit Mosteller und Wallace im Jahre 1964

ihr berühmtes Werk *Inference and disputed authorship, the Federalist*[49] veröffentlichten und nachdem ihre Ergebnisse durch die meisten nachfolgenden stilometrischen Arbeiten bestätigt wurden, kann Madison als der Autor aller 12 umstrittenen Artikel gelten.

Die einzelnen Artikel sind relativ lang. Dies und die Tatsache, dass nur zwischen zwei Autoren ausgewählt werden muss (John Jay stand nicht zur Debatte), macht die gestellte Aufgabe aus stilometrischer Sicht tendentiell eher einfach. Infolgedessen gibt es eigentlich nur die Arbeiten, die alle Artikel Madison zuschlagen und die, die das nicht vermögen. Eine weiter abgestufte Beurteilung der veröffentlichten Ergebnisse ist kaum möglich.[50]

Auf der anderen Seite hat sich das *Federalist*-Korpus zu einem gewissen Standard entwickelt. Eine neue Methode sollte diesen Test bestehen.

Es ist ein Vorteil des Datensatzes, dass *Topic* und *Genre* der Texte extrem homogen sind. Auch der Schreibstil der beiden in Frage kommenden Autoren Madison und Hamilton gilt als ausgesprochen ähnlich.[51]

Daneben gibt es aber noch einen anderen Grund, sich diesem Datensatz zu widmen: Er stellt gerade für die hier vorgestellte Methode einen interessanten Fall mit speziellen Problemen dar.

Die Literatur zu diesem speziellen Problem der Autorenbestimmung ist umfangreich. Adair (1944) wird meist als Referenzwerk für die Zeit vor der computerisierten Stilometrie zitiert, die mit Mosteller und Wallace (1964) begann. Seitdem waren die 85 Artikel Gegenstand zahlreicher Untersuchungen und auch in modernsten Arbeiten spielen sie noch eine Rolle (Jockers und Witten, 2010). Eine Übersicht würde an dieser Stelle keine über Abschnitt 3.2 hinausgehenden Erkenntnisse bringen.

Die Daten in ihrer hier verwendeten Form stammen von der Website des *Projekt Gutenberg* (Hamilton et al., 2004). Eine Zuordnung der einzelnen Texte zu ihren Autoren gibt Tabelle 3.5.

Als einziges S-Maß kommt in diesem Kapitel S_{log} zur Anwendung. Es hat sich bisher als das überlegene erwiesen und für einen weiteren Performanzvergleich ist der vorliegende Datensatz nicht geeignet.

Für die hier dargestellte Methode ergibt sich ein Problem daraus, dass wir nicht a priori wissen, ob alle umstrittenen Artikel denselben Autor haben, oder ob sich die Autorschaft beliebig auf Madison und Hamilton verteilt. Damit ist die bisher verwendete Normierung nicht mehr anwendbar. Man vergegenwärtigt sich dieses Problem leicht bei Betrachtung von Abbildung 3.3 (Seite 154). Angenommen, der erste Eindruck ist richtig und es wurden tatsächlich die Hälfte der Dokumente oder mehr von Autor Nummer 3 verfasst. Dann spiegelt die Dominanz der 3. Reihe die tatsächlichen Verhältnisse wider. Die verwendete Normalisierung mit Hilfe des Mittelwertes würde nun die tatsächlichen Verhältnisse verdecken und eine korrekte Klassifzierung verunmöglichen.

Für unseren Datensatz ist das Problem schematisch in Tabelle 3.6a dargestellt. Es wird so sein, dass die erste (Hamilton-)Zeile, generell höhere Werte enthält als die zweite

[49] Nicht *the Federalist Papers*, da der Originaltitel der Buchveröffentlichung lautete: *The Federalist: A collection of essays, written in favour of the new constitution, as agreed upon by the federal convention, September 17, 1787.*

[50] Es gibt zu dieser Frage natürlich auch die gegenteilige Meinung (Jockers und Witten, 2010, 3).

[51] Teahan (2000) zitiert Mosteller und Wallace (1984) zum Beleg.

Artikelnummer	Autor
1	Hamilton
2–5	Jay
6–9	Hamilton
10	Madison
11–13	Hamilton
14	Madison
15–17	Hamilton
18–20	Hamilton und Madison gemeinsam
21–36	Hamilton
37–48	Madison
49–58	**umstritten zwischen Madison und Hamilton**
59–61	Hamilton
62–63	**umstritten zwischen Madison und Hamilton**
64	Jay
65–85	Hamilton

Tabelle 3.5: Die *Federalist Papers*; Die Zuteilung zu den Autoren (oder eben nicht) folgt der traditionellen Einteilung wie sie zum Beispiel bei Holmes und Forsyth (1995) nachzulesen ist.

	d_1	d_2	d_3	...	d_{12}
T_H	$S(T_H, d_1)$	$S(T_H, d_2)$	$S(T_H, d_3)$	...	$S(T_H, d_{12})$
T_M	$S(T_M, d_1)$	$S(T_M, d_2)$	$S(T_M, d_3)$	...	$S(T_M, d_{12})$

(a)

	b_1	b_2	b_3	...	b_{100}
T_H	$S(T_H, b_1)$	$S(T_H, b_2)$	$S(T_H, b_3)$	...	$S(T_H, b_{100})$
T_M	$S(T_M, b_1)$	$S(T_M, b_2)$	$S(T_M, b_3)$	...	$S(T_M, b_{100})$

(b)

Tabelle 3.6: Schematische Übersicht über die zu berechnenden S-Werte. T_H bzw. T_M bezeichnen die Trainingskorpora aus bekanntermaßen von Hamilton bzw. Madison stammenden Texten. d_1 bis d_{12} bezeichnen die 12 umstrittenen Artikel. b_1 bis b_{100} sind die 100 BNC-Eichdateien. S steht hier immer für S_{log}.

(Madison-)Zeile. Dies schon deswegen, weil Hamilton viel mehr Aufsätze geschrieben hat und das Trainingskorpus somit länger ist. Falls nun alle umstrittenen Artikel tatsächlich von Hamilton geschrieben wurden, würden deswegen die 1. Zeile noch ein klein wenig höhere S-Werte enthalten. Dieser Effekt, für den wir uns ja eigentlich interessieren, würde durch die Normierung mit wegnormiert; die Normierung besteht ja darin, dass jeder S-Wert durch den Mittelwert der entsprechenden Zeile – und Spalte – geteilt wird.

Ich habe ein entsprechendes Experiment durchgeführt. Zwei Trainingstexte aus reinen

Hamilton- und reinen Madison-Dokumenten wurden mit jeweils 10 Testtexten ebenfalls bekannter Autoren verglichen. Ist das Testset ausgewogen zwischen Hamilton und Madison, so ergeben sich Erfolgsraten um die 97%. Dominiert ein Autor das Testset, so bricht die Qualität der Klassifikation schnell zusammen. Besteht es aus einem einzigen Autor, ergibt sich genau die Baseline von 50%.

Um dieses Problem zu lösen führe ich 100 Eichdateien b_i ein. Diese bestehen aus den jeweils ersten 35000 Zeichen willkürlich ausgewählter Dateien aus dem BNC (Burnard, Lou , Hg.). Mit deren Hilfe wird der Anteil an S berechnet, der nur von den Trainingstexten T_H und T_M abhängt. Nun wird der S-Wert für jedes Paar aus einer BNC-Eichdatei und einem Trainingstext berechnet. Es ergibt sich das in Tabelle 3.6b dargestellte Schema.

Aus diesen Werten lässt sich der vom Trainingskorpus T_H abhängige Anteil an S abschätzen durch den Mittelwert

$$S(T_H) = \frac{1}{100} \sum_{i=1}^{100} S(T_H, b_i)$$

Die nur von den Testdateien abhängigen Anteile von S-werden wie üblich abgeschätzt. Ich definiere den *BNC-geeichten Ähnlichkeitsindex* für die Dateien T_X und d_i:

$$S_{BNC}(T_X, d_i) = \frac{S(T_X, d_i)}{S(T_X)\frac{1}{3} \sum\limits_{Y \in \{H,M,J\}} S(T_Y, d_i)} \tag{3.3}$$

Hier laufen X und Y über H, M und J für die drei Autoren Hamilton, Madison und Jay. Jay ist hier noch nicht von Interesse, ich beziehe ihn aber von Anfang an in die Berechnung ein und komme später auf ihn zurück. Diese etwas komplexe Notation kann in einer einfachen Aussage zusammengefasst werden: In der Normierung werden die 12 umstrittenen Aufsätze durch die 100 BNC-Dateien ersetzt.

Auch für die übrigen, nicht umstrittenen Dateien wird ebenfalls ein *BNC*-geeichter Wert berechnet. Dazu wird das betrachtete Dokument aus dem entsprechenden Trainingskorpus entfernt (und behandelt als wäre seine Autorschaft umstritten). Auf diese Weise lässt sich die Verlässlichkeit der Methode überprüfen.

Alle 85 so berechneten S_{BNC}-Werte sind in Abbildung 3.18 dargestellt.

Die bekanntermaßen von einem der beiden Autoren Hamilton oder Madison geschriebenen Artikel lassen sich klar trennen. Nur zwei Artikel gehören jeweils „zur falschen Wolke“. Dies sind die Artikel 9 und 10. Sie liegen beide recht nah an der teilenden Linie, sind also noch als statistische Ausreißer zu erklären. Beide Artikel behandeln dasselbe (Unter)-Thema,[52] eine Verwechslung liegt also nahe, besonders da Artikel Nr. 10 einer von nur wenigen Artikeln ist, in denen ein Autor ein Thema des anderen fortführt. Die *cross validation* war also in 63 von 65 Fällen erfolgreich (97%).[53]

[52] „The Union as a Safeguard Against Domestic Faction and Insurrection“

[53] Andere Arbeiten berichten von Besonderheiten in Bezug auf andere Artikel. So finden sich in Teahan (2000) Bemerkungen zu den Artikeln 62 und 63, die zu den umstrittenen Artikeln gehören, sowie zu 59, der bei ihm fälschlich Madison zugeordnet wird. Mosteller und Wallace (1984) berichten von

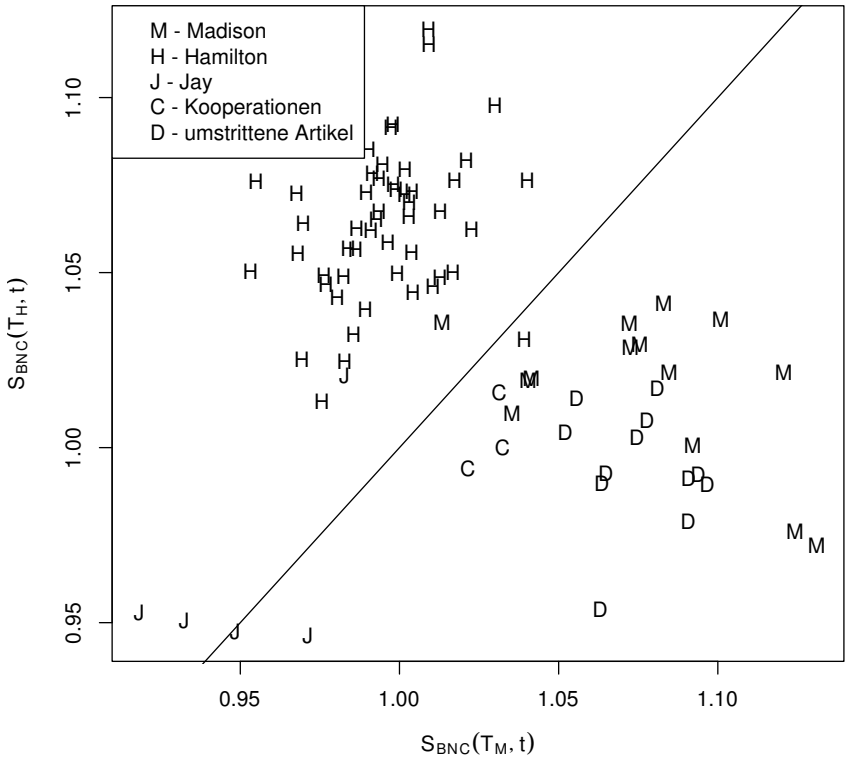


Abbildung 3.18: Die *BNC-geeichten* S-Werte. Die X-Achse bezeichnet die Ähnlichkeit zum Madison-Trainingstext, die Y-Achse die Ähnlichkeit zum Hamilton Trainingstext. Die Gerade ist die Diagonale $X = Y$.

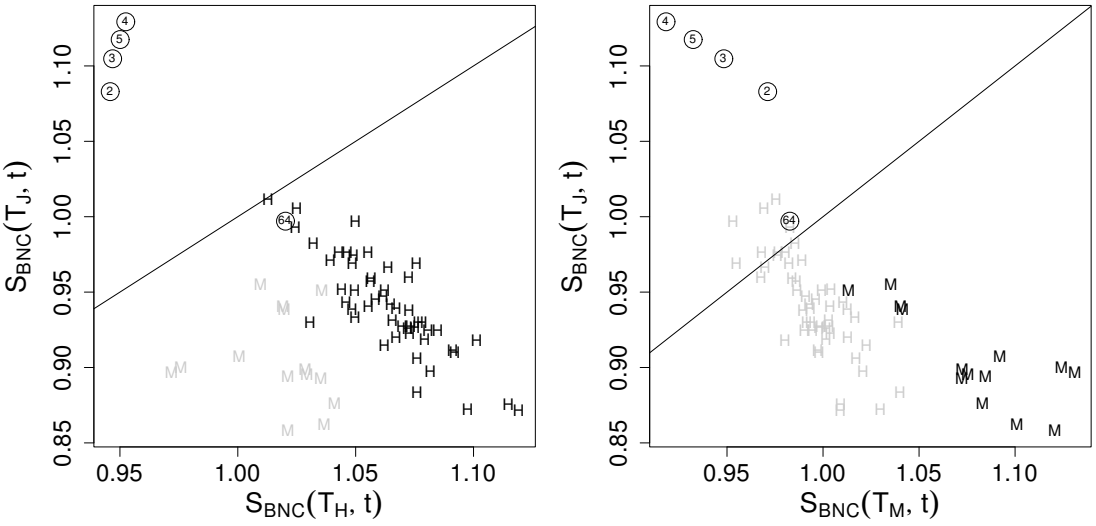
Die drei Kooperationen (im Bild bezeichnet mit „C“) liegen nahe der Trennungslinie, was nicht unplausibel scheint.

Die umstrittenen Artikel werden ausnahmslos Madison zugeschlagen. Nach dem Stand der Forschung kann dieses Ergebnis als etabliert gelten. Zusammen mit dem gelungenen Nachweis, dass die Methode auf den bekannten Texten zuverlässig funktioniert ist dies ein befriedigendes Ergebnis. Bemerkenswert an der verwendeten Methode ist, dass *amerikanische* Texte vom Ende des 18. Jahrhunderts mit Hilfe von um 200 Jahre jüngeren *britischen* Texten normalisiert werden konnten.

Der dritte Autor, John Jay, zeigt Besonderheiten. Vier der 5 Jay zugewiesenen Dateien haben geringe Ähnlichkeitswerte relativ zu den beiden anderen Autoren. Auch dies ist

Unklarheiten vor allem in Bezug auf Nummer 55. Holmes und Forsyth (1995) ordnen Artikel 70 fälschlicherweise Madison zu. Jockers und Witten (2010) erhält in seiner komparativen Studie Zuordnungsfehler für die Artikel 49, 50, 51, 52, 56 (jeweils 1 Fehler), 54 (2 Fehler), 55 und 57 (jeweils 3 Fehler)

wie erwartet. Der verbleibende Artikel ist Nummer 64. Ich habe Hinweise gefunden,[54] dass dieser Artikel nicht von Jay geschrieben sein könnte. Meine Ergebnisse weisen auch in diese Richtung, auch wenn ich ihn eher Hamilton als Madison zuschlagen müsste.



(a) (b)

Abbildung 3.19: *BNC-geeichte S*-Werte. Die Y-Achse bezeichnet in beiden Teilbildern die Ähnlichkeit zu Jay. Die X-Achse bezeichnet in Teilbild (a) die Ähnlichkeit zu Hamilton bezeichnet und in Teilbild (b) die Ähnlichkeit zu Madison. Die Gerade repräsentiert die Diagonale $X = Y$. Die mit Zahlen bezeichneten Artikel werden traditionell alle Jay zugeschlagen. Für 4 Artikel scheint das gerechtfertigt, Artikel 64 dagegen verhält sich abweichend.

Dies sieht schon in Abbildung 3.18 so aus und wird durch Abbildung 3.19 untermauert. Dort ist für dieselben Daten jeweils die Ähnlichkeit mit Jay (Y-Achse) mit der Ähnlichkeit eines der beiden anderen Autoren verglichen. Demnach scheint Jays Autorschaft unwahrscheinlich.

Sämtliche hier durch den bloßen Vergleich von S-Werten durchgeführten Klassifikationen lassen sich bestätigen, wenn man zur Klassifikation die in `R` verfügbaren Module zur Berechnung von *Support-Vektor-Maschinen* einsetzt. Nur, wenn man auf den zweiten Schritt der Normierung (Gleichung 3.3) verzichtet und alle 70 Dateien mit bekannten Autoren crossvalidiert, bekommt man 69 von 70 korrekt zugewiesen, alle bis auf Artikel 64. Aber dieses Ergebnis ist nicht allzu stabil.

In diesem Abschnitt wurde das entwickelte Verfahren auf ein altes Standardproblem der Stilometrie angewendet: *Automatische Autorenbestimmung* anhand der *Federalist*

[54] „No. 64 was by John Jay. Some newer evidence suggests James Madison as the author.“ (Wikipedia-Mitarbeiter, 2001), leider fehlt der Hinweis auf ein Belegstelle.

Papers. Die Methode erwies sich als erfolgreich in dem Sinne, dass die Ergebnisse der Forschungsgemeinde bestätigt werden konnten, die sich ganz überwiegend einig ist, dass alle 12 umstrittenen Aufsätze von James Madison verfasst wurden. Die Tatsache, dass nicht von vornherein bekannt ist wie sich die Autorenschaft der umstrittenen Texte verteilt, macht die Normierung in ihrer Grundform (Abschnitt 3.4) nutzlos. Die Einführung von 100 *Eichdateien* aus dem BNC (Burnard, Lou , Hg.) löst dieses Problem effizient, trotz des sprachlichen Abstands zum eigentlichen Testkorpus. Für zwei der nicht umstrittenen Artikel (#9 und #10) deutet sich die Möglichkeit einer *Topic*-Interferenz trotz der extremen Homogenität des Korpus an. Es ergeben sich Zweifel an der Autorschaft von John Jay an Aufsatz #64.

Nach einem wegen seiner Etabliertheit nützlichen Beispiel folgt nun eine stilometrische Untersuchung anhand eines völlig neuen Korpus und anhand einer völlig neuen Fragestellung. Die Struktur dieser Fragestellung ist so geartet, dass das Paradigma der Stilometrie als reiner Text*klassifikation* durchbrochen wird.

3.6.4 Hat S vererbbare Komponenten?

In diesem Abschnitt werden Untersuchungen anhand eines Korpus beschrieben, das von Mollet et al. (2010) im Rahmen eines größeren Projekts entwickelt wurde, dessen Ziel es ist, die genetische Komponente sprachlicher Variation zu untersuchen.

Die Hauptmotivation der Untersuchung liegt in der Klärung der Frage, ob sich für S (in der logarithmischen Variante S_{log}) erbliche Komponenten nachweisen lassen. In der gesamten stilometrischen Literatur ist mir kein zweites Beispiel für eine ähnliche Untersuchung bekannt. Auch der erwähnte Artikel von Mollet et al. ist eine Vorstudie mit einer wesentlich allgemeineren Fragestellung. Im Ergebnis wird sich ein augenfälliger Effekt zeigen, für den statistische Signifikanz allerdings nicht nachweisbar ist. Bereits dies aber ist angesichts des hoch komplexen Datensatzes ein berichtenswertes Ergebnis.

Auf dem Weg zu diesem Ziel wird uns der Datensatz Gelegenheit geben, die Verflechtung von *Genre* und *Topic* noch einmal genauer zu betrachten. Das *Genre* des Textes ist dabei vor allem als zweite bedeutende Einflussgröße neben dem *Topic* von Bedeutung. Die Ausführungen in Abschnitt 3.2 auf Seite 135 motivieren ihren Einfluss. Hier zeige ich nun, dass *Topic* und *Genre* nicht nur beide großen Einfluss auf S haben, sondern dass dieser Einfluss auf verschiedenen Ebenen des Textes angesiedelt ist. Grob gesagt kann das *Topic* aus den Inhaltswörtern des Textes abgelesen werden, während das *Genre* eher in den Funktionswörtern steckt. Daher besteht die Gefahr, dass naive Filterungen des Textes zur *Topic*- oder *Genre*-Elimination in Bezug auf die jeweils andere Variable einen verstärkenden Einfluss haben. Dies wäre für weite Teile der stilometrischen Forschung ein interessantes Ergebnis. Wie in Abschnitt 3.2 dargestellt sind die Funktionswörter ihre verbreitetste Datenquelle. Ein Grund hierfür ist die Bestrebung, so den Einfluss des *Topics* zu neutralisieren. Wenn dies wiederum den *Genre*-Effekt stärkt, muss im mindesten auf eine erhöhte *Genre*-Homogenität geachtet werden. Sonst ist die Gefahr gegeben, dass als stilometrisch bewertete Klassifikations(miss)erfolge in Wahrheit auf *Genre*-Inhomogenitäten zurückgehen.

In den Untersuchungen dieses Abschnitts wird auch wieder der Einfluss der Repräsen-

tation des Textes eine Rolle spielen, indem die inhaltstragenden Wörter durch ihre POS-Tags ersetzt werden. Im Lichte des hier untersuchten, vollkommen unterschiedlichen Datensatzes zeigen sich durchaus neue Aspekte gegenüber den Erkenntnissen aus der Analyse des *Translationese*-Korpus in Abschnitt 3.6.1.

Auch das Thema der Korpusreinheit wird noch einmal aufgegriffen. Hier geht es insbesondere um das Problem durch Verunreinigungen infolge direkter Textübernahmen. Dieses Phänomen ergibt sich beinahe zwangsläufig, wenn Texte untersucht werden, die aufgrund direkter thematischer Vorgaben geschrieben werden (elizitierte Daten). Viele korpuslinguistische Korpora, va. auf dem Gebiet der Lernersprachenforschung sind von einer derartigen Struktur. Wir sind bereits beim ICLE in Abschnitt 3.6.2 auf das Phänomen und die daraus erwachsenden Probleme gestoßen. Hier werde ich eine sensiblere Methode, mit dem Problem umzugehen, vorstellen.

Abgerundet werden der Abschnitt durch Untersuchungen zu alternativen Formen der Normalisierung von S. Bisher wurde nur die sehr einfache heuristische Form der Normalisierung betrachtet wie sie in Abschnitt 3.4 und insbesondere mit Gleichung 3.2 beschrieben wurde. Hier werden nun analytischere Formen betrachtet und die damit verbundenen Verbesserungen diskutiert.

Als Erblichkeit wird der Anteil an der Variabilität einer Eigenschaft bezeichnet, die auf genetische Unterschiede zurückzuführen ist. Um diesen messen oder abschätzen zu können gilt es, den Einfluss der Gene vom Einfluss der Umgebung zu trennen. Eine Methode hierfür sind Zwillingsstudien. Deren Ausgangspunkt ist, dass es zwei Arten von Zwillingen gibt, eineiige (monozygotisch, MZ) und zweieiige (dizygotisch, DZ). Das Erbmaterial von eineiigen Zwillingen ist beinahe vollständig identisch, während zweieiige Zwillinge so viel Erbgut teilen wie normale Geschwister. Die Grundannahme der Zwillingsforschung ist, dass die Umgebung beider Arten von Zwillingen als identisch angenommen werden kann. Verglichen wird nun die Ähnlichkeit von eineiigen Zwillingen mit der von zweieiigen. Erweisen sich eineiige Zwillinge in Bezug auf die untersuchte Eigenschaft ähnlicher als zweieiige, wird angenommen, dass nicht nur die Umgebung, sondern auch die Gene einen Einfluss haben.

Quantitativ wird die Erblichkeit über die Differenz der beiden Fälle abgeschätzt. Mit den Gleichungen, die konkret verwendet werden, müssen wir uns hier nicht beschäftigen, da die benötigten Korrelation mit den gegenwärtig zur Verfügung stehenden Mitteln nicht berechnet werden können. Es ist aber möglich zu untersuchen, ob es überhaupt einen Unterschied in der Ähnlichkeit der beiden Zwillingsvarianten gibt. In diesem Sinne ist die vorliegende Untersuchung qualitativ.

Das Korpus besteht aus 55 Aufsätzen. Jeder Aufsatz ist von einem anderen Autor. Alle sind auf Englisch[55], der Muttersprache der Autoren. Diese teilen sich in 5 eineiige Zwillingspaare, ein zweieiiges Drillingspaar und 21 zweieiige Zwillingspaare[56]. Alle waren zum Zeitpunkt der Erhebung 17 Jahre alt. 22 der Versuchspersonen waren männlich, 33 weiblich. Das Genre der Aufsätze war den Versuchspersonen überlassen. Die Verteilung der Genres ist in Tabelle 3.7 aufgelistet. Die Länge der Texte liegt relativ gleichverteilt

[55] Australisches Englisch
[56] $5 \cdot 2 + 3 + 21 \cdot 2 = 55$

zwischen 456 und 940 Wörtern.[57]

Genre	Anzahl
Erörterung	16
Erzählung in der 1. Person	10
Erzählung in der 3. Person	8
Introspektion	6
Brief	5
Rede	3
Feature	3
Tagebucheintrag	2
Drehbuch	1
Literaturkritik	1

Tabelle 3.7: Übersicht über die im Korpus (Mollet et al., 2010) vertretenen Genres, geordnet nach Häufigkeit. Die Zuordnungen stammen von den Autoren des Korpus.

23 der Aufsätze wurden 2004 geschrieben, 32 stammen von 2005. In jedem der beiden Jahre war ein unterschiedliches Thema vorgegeben. Dieses wurde nicht durch einen einzigen Satz bezeichnet wie bei der Erhebung des ICLE-Korpus (Granger, 2003). Statt dessen wurde den Probanden relativ umfangreiches Stimulusmaterialien vorgelegt, bestehend aus Text und Bildern. Der Text umfasste 426 bzw 727 Token. Die Zwillings- und Drillingspaare haben ihren Aufsatz immer im jeweils selben Jahr geschrieben, und damit auch zum selben Topic.

Das Korpus stellt stilometrisch eine sehr schwierige Aufgabe dar: Stilistische Einflüsse auf Oberflächenphänomene sind relativ schwach verglichen mit dem Einfluss des *Topics* (Golcher und Reznicek, 2011). In den bisher untersuchten Korpora spielte das keine allzu große Rolle. Im *Translationese*-Korpus (Abschnitt 3.6.1) hatte jeder Text sein eigenes Thema und die Themen alle aus einer einzigen Domäne, der Geopolitik. Im ICLE-Korpus (Abschnitt 3.6.2) gibt es sehr viele unterschiedliche *Topics*, die sich zugleich wiederholen. Insgesamt weisen sie aber eine so breite Streuung auf, dass man aufgrund der hohen Zahl der Texte hoffen kann, dass sich ein verfremdender Effekt weitgehend herausmittelt. Die *federalist papers* hingegen sind thematisch sogar noch homogener als das *Translationese*-Korpus, sie befassen sich alle mit der amerikanischen Verfassung. Im jetzt zu untersuchenden Korpus gibt es dagegen zwei stark unterschiedliche Themen, die sich darüber hinaus mit der Einteilung in Zwillingspaare decken. Dies macht es sehr schwer, den genetischen Einfluss vom Einfluss des *Topics* zu trennen.

Wie das *Topic* wird auch das *Genre* einen Einfluss auf S haben, bzw. die relative Ähnlichkeit der *Genres* der beiden verglichenen Texte. Hinzu kommt, dass die von den Autoren des Korpus gegebenen Genrebezeichnungen (Tabelle 3.7) keine sehr strenge Klassifizierung darstellen.

Das Korpus widerspricht damit den etablierten Regeln für ein Korpus dieser Art:

[57] Geschätzt mit dem Standardtool `wc` unter linux.

„Any good evaluation corpus for authorship attribution should be controlled for genre and topic“ (Stamatatos, 2009, 552). Da beide Variablen nicht kontrolliert sind, sondern stark und frei variieren, müssen sie explizit in die Berechnung mit einbezogegen werden.

Von der vermutlich recht starken Störung durch *Topic* und *Genre* gilt es die stilistische Einflüsse aber nicht nur als solche abzuspalten. Darüber hinaus soll nach Möglichkeit gezeigt werden, dass ein Teil der Autoren untereinander ähnlicher schreibt als ein anderer. Dafür stehen mit nur 5 eineiigen Zwillingspaaren und nur einer halben bis ganzen Seite Text pro Autor ausgesprochen wenig Daten zur Verfügung. Es gibt aber noch ein weiteres Problem, das im Rahmen der Untersuchung des ICLE-Korpus (Abschnitt 3.6.2) bereits (in abgeschwächter Form) auftritt:

Die Tatsache, dass die Probanden Stimulusmaterial präsentiert bekamen, anhand dessen sie ihre Texte schrieben, hat neben dem thematischen Bias an sich noch eine weitere Folge. Das Stimulusmaterial findet über mehr oder minder lange direkte Zitate Eingang in die Aufsätze. Dies wäre an sich schon ein Problem, da man annehmen kann, dass derartige Übernahmen das stilistische Signal in schwer abschätzbarem Ausmaß beeinflussen. Hinzu kommt aber, dass so auch identische Sätze den Weg in verschiedene Aufsätze finden. Dies führt unmittelbar zu extrem hohen $S(T_i, T_j)$-Werten zwischen den entsprechenden Aufsätzen. Daher ist es erforderlich, die Texte von diesen unnatürlichen Wiederholungen zu reinigen. (Unnatürlich in dem Sinne, dass die Texte keine so langen Zeichenketten teilen würden, wenn die Verfasser nicht dasselbe Stimulusmaterial vor sich gehabt hätten.)

Hier, wo der Stimulus selbst längere Texte beinhaltet, kann erwartet werden, dass das Problem eher noch stärker in Erscheinung tritt als beim ICLE-Korpus, wo die Studenten nur einzelne Sätze vorgegeben bekamen. Dort wurden Wiederholungen der Themenstellung mit einer relativ einfachen Heuristik aus den Lernertexten entfernt. Alleine weil hier das Stimulusmaterial viel umfangreicher ist, wird eine Filterung nicht so einfach möglich sein.

Was also soll als „unnatürliche Wiederholung“ zählen? Folgende Heuristik wird verwendet, um unnatürliche Wiederholungen zu identifizieren: Grundsätzlich werden alle Zeichenketten aus den Aufsätzen entfernt, die genau einmal auch im Stimulanzpapier vorkommen. Da dies leicht zu zufälligen Treffern führt, werden nur die Strings entfernt, bei denen dieses Kriterium auch dann zutrifft, wenn vorne und hinten einige Zeichen entfernt werden.

Der Text des Stimulusmaterials, der *Quelltext* sei mit T_s bezeichnet. Der zu filternde Aufsatz heiße A. Zur Identifizierung von *kopiertem Material* verwende ich folgende Definition:

Definition 38 (kopiertes Material der Ordnung n) *Eine Zeichenkette s des Testtextes A, die genau einmal im Quelltext T_S vorkommt, heißt kopiertes Material der Ordnung n, wenn s hinten und vorne um n Zeichen gekürzt werden kann, so dass die gekürzte Version immer noch genau einmal in T_s vorkommt.*

Die Definition hat den Zweck unnatürliche, nur durch den Wortlaut des Stimulusmaterials verursachte, Wiederholungen zu identifizieren, aber gewöhnliche oder *topic-*

typische Wiederholungen zu akzeptieren: Ein Text über „Rosenzucht“ wird das Wort Rose wahrscheinlich mehrfach enthalten, unabhängig davon wie die genaue Fragestellung formuliert ist. Derartige Wiederholungen stufe ich als normal ein. Sie sollen erhalten bleiben. Bei Zeichenketten, die sich im *Quelltext* bereits wiederholen, wird angenommen, dass es sich um derartige thementypische Wiederholungen handelt.

Nun sei anhand eines Beispiels demonstriert wieso es nötig ist, einzigartige Wiederholungen an den Enden zu beschneiden, um zufällige Wiederholungen herauszufiltern:

Quelltext: `Do_we_have_beer_or_do_we_have_wine,_Josef?`

Testtext: `Someone_must_have_been_telling_lies_about_Josef_K.`

Sei $n = 1$. Dann gilt `_Josef` als *kopiertes Material*, weil nicht nur diese Zeichenkette selber nur einmal in T_s erscheint, sondern auch die verstümmelte Version `Jose`. Die Zeichenkette `_have_b` dagegen, die ja auch nur einmal in T_s auftaucht, gilt nicht als kopiertes Material. Ohne den ersten und letzten Buchstaben bekommen wir hier `have_`. Diese Zeichenkette allerdings taucht zwei Mal im *Quelltext* auf und fällt somit nicht unter obige Definition.

Neben diesem Problem direkter Übernahmen aus dem *Quelltext* bleibt aber nach wie vor der *Topic*-Effekt selbst bestehen, der davon unabhängig ist. *Topic*-basierte Wiederholungen können leider nicht aus dem Text selbst herausgeschnitten werden wie direkte Quelltextübernahmen.

Wie im Forschungsüberblick 3.2 bereits angesprochen, gibt es die verbreitete Überzeugung, dass Funktionswörter oder POS-Tags frei von thematischen Einflüssen sind. Anders herum ausgedrückt, dass das Entfernen der Oberflächenformen der inhaltstragenden Wörter oder ihr Ersetzen durch POS-Tags den thematischen Einfluss eliminiert.

Diese Annahme soll hier empirisch überprüft werden. Einerseits gibt das Hinweise darauf, ob dieses allgemein verwendete Verfahren tatsächlich angemessen ist. Anderseits kann so entschieden werden, ob das Entfernen lexikalischen Inhalts im vorliegenden Fall eine Option ist. Für diese Untersuchung habe ich eine Korpusrepräsentation erstellt, in der alle inhaltstragenden Wörter durch die entsprechenden POS-Tags ersetzt wurden. Die Annotierung erfolgte mit Hilfe des Treetaggers von Schmid (1994). Diese Variante der Texte entspricht im Wesentlichen der `mix`-Darstellung, die schon in Abschnitt 3.6 in Anlehnung an Baroni und Bernardini (2006) untersucht wurde. Hier nenne ich sie die *inhaltsleere* Darstellung des Korpus.

Die Wirkung verschiedener Filterstufen n auf die beiden Korpusrepräsentationen ist in Abbildung 3.20 zusammengefasst. Es ist tatsächlich so, dass das Ersetzen der Inhaltswörter durch POS-Tags den *Topic*-Effekt stark reduziert. Für *Genre* zeigt sich ein gegenläufiger Effekt: Ohne Inhaltswörter tritt der Unterschied der Genres erheblich stärker zu Tage. Es ist durchaus plausibel, dass das *Topic* sich am lexikalischen Inhalt zeigt, während das *Genre* eher durch strukturelle Eigenheiten beschrieben werden kann. Schon aus diesem ersten Blick auf die Graphik kann die Konsequenz gezogen werden, dass die Ersetzung der Oberflächenwortformen durch POS-Tags vor allem im vorliegenden Fall nicht anzuraten ist. Auf der Gewinnseite stünde zwar das Unterdrücken des Topic-Effektes. Dem steht aber eine erhebliche Verstärkung des Genre-Effektes gegenüber.

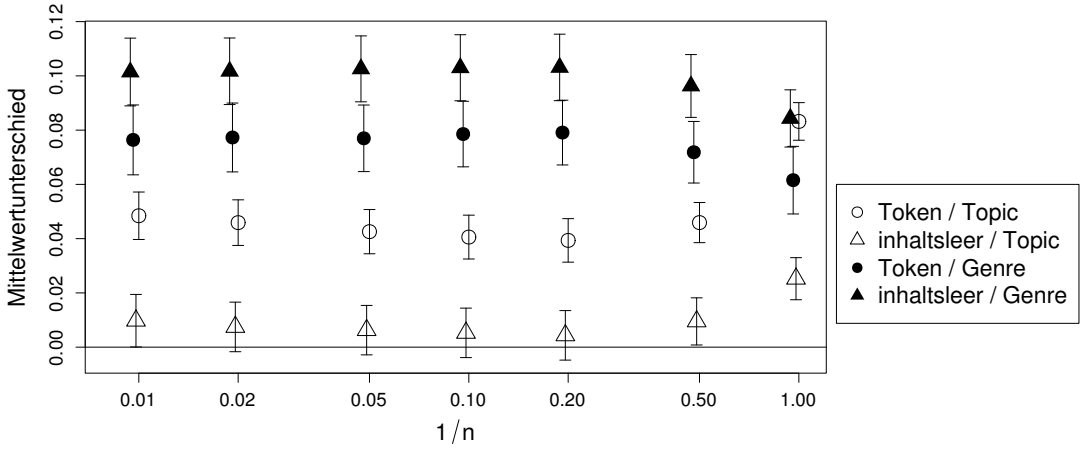


Abbildung 3.20: Der Einfluss des Entfernens von Inhaltswörtern auf die Unterscheidbarkeit der Einflüsse von Genre und Stimulusmaterial (*Topic*). Die X-Achse stellt den Kehrwert der *Ordnung* n dar. Dh., bei $x = 0$ liegt keine Filterung vor, da $n = \infty$. Bei $x = 1$ dagegen wird stark gefiltert, da hier auch $n = 1$ ist. Für die Y-Achse wurden die $S(T_1, T_2)$-Werte jeweils in zwei Gruppen aufgeteilt, je nachdem ob T_1 und T_2 das *Genre* oder das Stimulusmaterial teilen, oder nicht. Der Unterschied der Mittelwerte ist der dargestellte Wert. Die ausgefüllten Punkte bezeichnen die Unterschiede in Bezug auf das *Genre*, die leeren Punkte in Bezug auf das Stimulusmaterial (*Topic*). Die Kreise stehen jeweils für den originalen Text. Für die Dreiecke wurden die Inhaltswörter durch POS-Tags ersetzt. Die Fehlerbalken zeigen das Konfidenzintervall eines t-Tests auf Mittelwertgleichheit an.

Dies wäre wegen der starken Genre-Inhomogenität des Korpus stark nachteilig. Darüber hinaus ist eine Ersetzung allen lexikalischen Inhalts mit einem großen Verlust an Information verbunden. Dass auch lexikalische Information für stilometrische Fragestellungen eine Rolle spielen kann, wurde in Abschnitt 3.6.1 gezeigt.

Das gegenläufige Verhalten von *Topic* und *Genre* wird leicht ignoriert oder nicht verstanden von Autoren, die den Einfluss des *Topics* auf ihren stilometrischen Ansatz durch die Beschränkung auf Funktionswörter umgehen zu wollen.

So vermischen Clement und Sharp (2003) die beiden Variablen, wenn sie einerseits schreiben: „[...] genre-removing transformations (such as replacement of words with part of speech tokens)[...]“ um ein wenig später zu ergänzen „In any case, it appears that tagged versions of documents have lost all topic signal [...]“.[58]

Nicht nur in Bezug auf die Ersetzung von Oberflächenwortformen durch POS-Tags

[58] Erschwerend kommt für ihren Datensatz hinzu, dass die Autoren teils Briten und teils Amerikaner waren. So ist nicht auszuschließen, dass der persönliche Stil eines Autors mit seiner Sprachvariante vermischt wird. Clement und Sharp (2003) selbst erwähnen die Neigung mancher der untersuchten Autoren zu bestimmten Themen, ein Umstand, der die Deutung der Ergebnisse weiter verkompliziert.

verhalten sich *Topic* und *Genre* gegenläufig. Ein ähnliches Verhalten beobachten wir in der Entwicklung in Abhängigkeit von der *Ordnung* n. Auf der X-Achse ist nicht direkt n aufgetragen, sondern der Kehrwert $1/n$. Diese Darstellung schien mir intuitiver, da nun das Ausmaß an Filterung von links nach rechts zunimmt.

Für eine sehr geringe Filterung[59] ist der *Topic*-Effekt in der *inhaltsleeren* Darstellung gerade eben signifikant. In der Originaldarstellung ist er ohnehin klar erkennbar. Mit stärker werdender Filterung nimmt er in beiden Darstellungen ab und erreicht bei $1/n = 0.2$ bzw $n = 5$ ein Minimum. Nun ist für die *inhaltsleere Darstellung* kein *Topic*-Effekt mehr nachweisbar. Der *Genre*-Effekt entwickelt sich gegenläufig: Mit steigender Filterung nimmt er langsam zu, bis er bei $n = 5$ ein (sehr flaches) Maximum ausbildet.

Dieses Verhalten suggeriert, dass der messbare *Topic*-Effekt in der POS-Darstellung des ungefilterten Textes auf direkte Übernahmen aus dem Stimulusmaterial zurückgeht.[60] Bei $n = 5$ ist der Einfluss des *Topics* minimal, während die Struktur des Textes noch intakt ist. Dies ist daran erkennbar, dass hier der *Genre*-Effekt sein Maximum hat. Für noch stärkere Filterung dagegen geht der Einfluss des *Genres* stark zurück, was auf ein Auflösen des Textzusammenhangs hindeutet.

Da sich hoffen lässt, dass nicht nur das *Genre*-Signal hier sein Optimum hat, sondern auch das uns interessierende stilometrische Signal, arbeite ich im weiteren Verlauf mit einer *Ordnung* von 5. Bemerkenswert ist das Übereinstimmen dieses Optimums mit den in Golcher und Reznicek (2011) berichteten Werten. Dies deutet darauf hin, dass es im Wesentlichen unabhängig von der Sprache und von der genauen Aufgabenstellung ziemlich eindeutig festlegbar ist, was eine *unnatürliche* Wiederholung ist.

Die minimale Erhöhung des Genre-Effektes durch die Filterung stellt weniger ein Problem dar, sondern deutet vielmehr auf eine Erhöhung der Auflösung der Methode durch die Reiningung des Korpus hin. Der Einfluss des Genres wird allerdings ausgeglichen oder herausgerechnet werden müssen. Der Topic-Effekt erfordert im vorliegenden Fall eine besondere Behandlung, da er mit der Einteilung der Zwillingspaare verflochten ist.

Es sei daran erinnert, dass auf das Ersetzen der Oberflächenwortformen durch POS-Tags verzichtet wurde. Datengrundlage bildet die Repräsentation des Textes, die in Abbildung 3.20 durch den offenen und den ausgefüllten Kreis bei $1/n = 0.2$ repräsentiert wird.

Ich komme nun zur Untersuchung der Fragestellung inwieweit Verwandtschaftsbeziehungen aus dem Korpus extrahiert werden können. Hier stellen sich weitere methodische Probleme. Der Begriff der Vererbbarkeit wie oben dargestellt baut auf dem Konzept eines messbaren Merkmals auf. In diesem Sinne ist aber *Stil* nicht messbar. S und seine Varianten messen nicht direkt *Stil*, sondern nur die Ähnlichkeit von Texten. Selbst wenn es gelingt, die Einflüsse von *Topic* und *Genre* herauszufiltern, bleibt immer noch lediglich eine Aussage über die stilistische Ähnlichkeit, keine Quantifizierung des *Stils* selbst.

Es liegt nahe, die Ähnlichkeit der Texte mittels eines der etablierten Clustering-

[59] $n = 100$ bedeutet, dass Wiederholungen mindestens 200 Zeichen lang sein müssen, um als *kopiertes Material* herausgefiltert zu werden.

[60] In Golcher und Reznicek (2011) wird gezeigt, dass sich dieses Phänomen auch in einem ganz anders strukturierten Korpus zeigt. Dies deutet auf einen verallgemeinerbaren Effekt hin. Allerdings kann das Filtern dort den *Topic*-Effekt nicht ganz aus der POS-Darstellung entfernen.

Verfahren zu Gruppen zusammenzufassen, in der Hoffnung dadurch die Verwandtschaftsverhältnisse abzubilden. Dazu wäre nur die mit S gemessene Ähnlichkeit in etwas mit einer Entfernung vergleichbares umzurechnen. Dies könnte der Kehrwert von S sein, da so Textpaaren großer Ähnlichkeit kleine Entfernungen zugeordnet würden. Es ist aber unvermeidlich, dass der Einfluss der Verwandtschaft gegenüber den starken *Topic*- und *Genre*-Effekten schwer erkennbar bleibt. Auch wären Signifikanzrechnungen schwer durchzuführen.

Eine weitere Möglichkeit, die es abzuwägen gilt, sind die Modelle und Verfahren, die in 2.6 eingesetzt wurden. Dies würde auf eine direkte Modellierung von S über ein lineares Modell hinauslaufen, in das u.a. *Topic* und *Genre* beider Texte eingehen würden. Ein Einwand gegen dieses Vorgehen ist allerdings die besondere Verteilung von S, wie sie in Abbildung 3.10 (Seite 168) zu sehen ist. Dort ist zu erkennen, dass der Großteil der Daten zwar annähernd normalverteilt ist. Die Verteilung der Daten läuft aber nach rechts in einen sehr langen flachen Schwanz aus. Diese Werte als Ausreißer auszuschließen ist nicht angemessen, da dieses Verteilungsbeispiel durchaus charakteristisch für die Verteilung von S-Werten ist. Da zu erwarten ist, dass der gesuchte Verwandtschaftsanteil relativ schwach ist, vor allem im Vergleich zum Einfluss von *Topic* und *Genre*, scheinen allzu weitgehende Näherungen in diesem Fall nicht hilfreich zu sein. Aus diesen Gründen habe ich mich für ein parameterfreies Verfahren entschieden, das stichhaltige Schlussfolgerungen zulässt.

Zu jedem der 55 Texte T_i werden alle $S(T_i, T_j)$ mit $i \neq j$ betrachtet. Diese S-Werte werden absteigend geordnet. Würde Verwandtschaft einen starken Einfluss auf die Textähnlichkeit haben und gäbe es keine weiteren Einflussgrößen, so könnte man erwarten, dass jeweils dasjenige $S(T_i, T_j)$ an erster Stelle steht, für das die Autoren i und j Zwillinge sind. Für jeden Text wird nun der tatsächliche Rangplatz r des Zwillingstextes registriert. Im Falle der Drillinge werden die Rangplätze beider Geschwister notiert.

Wendet man diese Methode auf die rohen, unnormalisierten S_{log}-Werte an, so bekommt man einen mittleren Rangplatz von $\overline{r} = 21.62$, was schon über der Baseline von $(n - 1)/2 = 54/2 = 27$ liegt. Für die eineiigen Zwillinge erhält man $\overline{r}_{MZ} = 14.9$ und für die zweieiigen $\overline{r}_{DZ} = 23.02$. Bereits dieses Ergebnis lässt darauf schließen, dass sich eineiige Zwillinge wesentlich leichter identifizieren lassen als zweieiige. Dies ist in unserem Sinne, wenn wir in S erbliche Faktoren nachweisen wollen.

Diese und die folgenden Rangplatzergebnisse sind in Tabelle 3.8 aufgelistet und in der dazugehörigen Graphik visualisiert.

Verwendet man die in Abschnitt 3.4 mit Gleichung 3.2 eingeführte Normierung, so bekommt man bereits bessere Ergebnisse. Der mittlere Rangplatz liegt nun bei $\overline{r} = 18.84$. Die Verbesserung kommt hier vor allem von den zweieiigen Zwillingen: $\overline{r}_{MZ} = 15.3$ bzw. $\overline{r}_{DZ} = 19.58$.

Diese Normierung beruht auf der Beobachtung, dass $S(T_i, T_j)$ für eine Menge an Texten T_i, $i \leq n$ nicht nur von der Ähnlichkeit der Texte alleine abhängt, sondern auch Beiträge enthält, die spezifisch für die eingehenden Einzeltexte sind. Diese wurden durch eine einfache Mittelwertbildung ausgeglichen.

Die Tatsache, dass hier jeder der Texte des Korpus mit allen anderen verglichen wurde, erlaubt es, über diese Heuristik hinauszugehen. Ich stelle ein explizites Mod-

ell für $S(T_i, T_j)$ auf, aus dem sich eine alternative Form der Normierung ableiten lässt. Ein multiplikatives Modell ist ein vielversprechender Ausgangspunkt:

$$S(T_i, t_j) = S_i S_j S_{ij} + \epsilon_{ij} \tag{3.4}$$

wobei S_i und S_j die nur von einem Text abhängigen Anteile bezeichnen. S_{ij} dagegen ist der Teil von S, der die Ähnlichkeit beider Texte beinhaltet. Ihm gilt unser Hauptinteresse. ϵ ist ein Fehlerterm unbekannter Verteilung mit $E(\epsilon) = 0$. Da es nur $n(n-1)/2$ unterschiedliche Messwerte für $S(T_i, T_j)$ gibt, das Modell aber selbst ohne den Fehlerterm n Parameter mehr enthält[61] sind nicht alle seine Parameter eindeutig bestimmbar. Ohne Einschränkung kann angenommen werden, dass $\overline{S_{ij}} = 1$.

Es wäre hilfreich, die Beiträge S_i zu eliminieren. Nehmen wir für den Moment an, dass sich der Fehlerterm vernachlässigen lässt. Umformen der Modellgleichungen ergibt dann folgenden Ausdruck für S_i:

$$S_i = \sqrt{\frac{S(T_p, T_i)\, S_{pq}\, S(T_q, T_i)}{S_{pi}\, S(T_p, T_q)\, S_{qi}}} \quad ; \quad p,\, q,\, i \text{ paarweise verschieden} \tag{3.5}$$

Die Werte für S_{pq}, S_{pi} und S_{qi} sind unbekannt. Insgesamt ergeben sich $(n-1)(n-2)/2$ Gleichungen für die verschiedenen Kombinationen für p und q. Da die S_{ij} als um 1 verteilt angenommen wurden ist es eine vernünftige Annahme, dass auch der Term $\sqrt{\frac{S_{pq}}{S_{pi} S_{qi}}}$ um 1 verteilt ist. Damit ergäbe eine Mittelung über den anderen Teil von Gleichung 3.5

$$S'_i = \sqrt{\frac{S(T_p, T_i) S(T_q, T_i)}{S(T_p, T_q)}} \tag{3.6}$$

eine Schätzung für S_i. Die entstehenden Werte kann man wiederum verwenden, um aus den $S(T_i, T_j)$ mit Hilfe der Ausgangsgleichungen 3.4 auf die S_{ij} zurückzuschließen. Diese Werte ersetzen das bisher verwendete S_{norm}. Die Unterschiede der Klassifikationsqualität, die sich bei Übergang von der bisherigen einfachen Normierung zu der soeben beschriebenen Form ergeben, sind in Abbildung 3.21 dargestellt. Die neue Form der Normierung führt zwar in 4 Fällen zu einer Verschlechterung des Rangplatzes um einen Punkt und aber in 11 Fällen zu einer Verbesserung von bis zu vier Rangplätzen. Es sei daran erinnert, dass mit dem Rang eines Textes die Zahl der Texte bezeichnet wird, die ihm ähnlicher scheint als der Text seines Zwillings. Der vorteil von 11 Verbesserungen gegenüber 4 Verschlechterungen ist statistisch allerdings nicht signifikant. So ergibt ein Binomialtest einen p-Wert von 0.12 für eine Abweichung von einer 50% Chance für positive wie negative Veränderungen. Ein ebenfalls durchgeführter Wilcoxon Rangsummentest, der die unterschiedlche Größe der Veränderungen einbezieht, ergibt einen p-Wert von 0.0508, ist also ebenfalls nicht signifikant. Man kann aus den Daten also die Hypothese ableiten, dass die genauere Normalisierung auch bessere Ergebnisse liefert, nachweisen kann man das mit den vorliegenden Daten nicht überzeugend.[62] Andersherum

[61] $n(n-1)/2$ für die $S_{i,j}$ und n für die S_j.

[62] Auch ein exakter Wilcoxon-Test auf Grundlage der genauen Rangverschiebungen (Hothorn und Hornik,

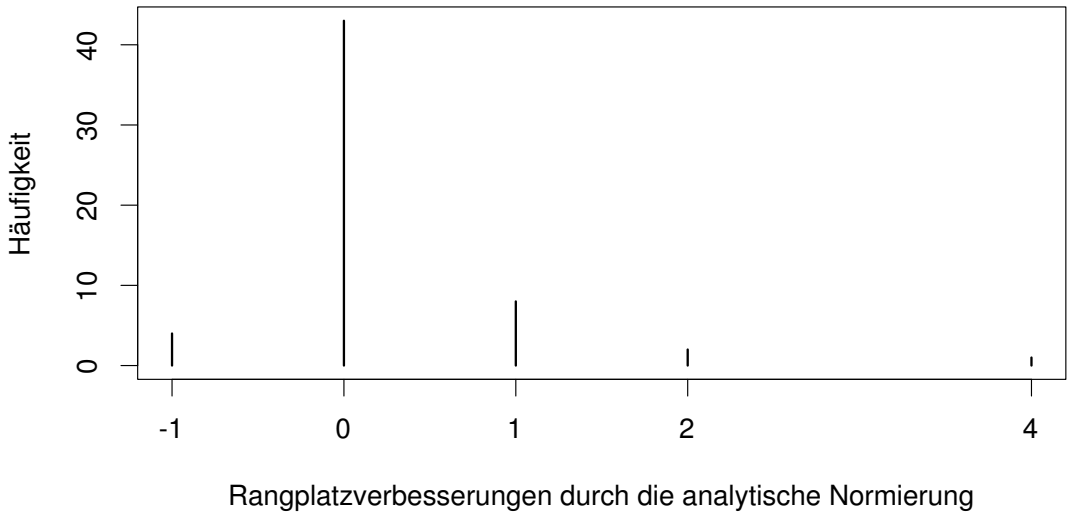


Abbildung 3.21: Rangplatzverschiebungen, die sich mit der analytischen Form der Normierung ergeben. Die positiven Verschiebungen überwiegen zwar, allerdings ist ihr Übergewicht nicht signifikant.

betrachtet heißt das, dass das bisherige Verfahren, S_i relativ grob zu schätzen, sich kaum nachteilig ausgewirkt haben dürfte.

Eine ähnliche Untersuchung wurde mit einem leicht veränderten Modell ausgeführt. Ein multiplikatives Modell scheint zwar nicht unplausibel, da durch Null nach unten begrenzte Größen eine Neigung zeigen, erst unter logarithmischer Transformation symmetrische Verteilungen anzunehmen. Beweisen lässt sich die Angemessenheit des Modells (3.4) so nicht. Auch ein additives Modell wie

$$S(T_i, t_j) = S_i + S_j + S_{ij} + \epsilon_{ij} \tag{3.7}$$

würde das in den Matrixplots wie (z.B. Abbildung 3.3) zu beobachtende Streifenmuster erklären. Dies führt zu einer entsprechenden Anpassung von Gleichung 3.5. Diese Normierung bringt allerdings keine signifikanten Vorteile.

Haben zwei Autoren sich für dasselbe Genre entschieden, so erhöht sich dadurch das S ihrer beiden Texte nicht unerheblich. Dies geht bereits aus Abbildung 3.20 hervor. Nachdem die Ergebnisse durch eine sorgfältigere Normierung bereits verbessert werden konnten soll nun überprüft werden, ob der Genre-Effekt herausgerechnet werden kann, um so den mittleren Rangplatz weiter zu erhöhen. Dafür wurden aus allen $S(T_i, T_j)$-Werten zwei Gruppen gebildet, je nachdem, ob T_i und T_j zum selben Genre gehören oder nicht. Anschließend wurde jeder Wert durch den Mittelpunkt seiner Gruppe geteilt. Bereits dieses sehr simple Vorgehen führt zu einer erheblichen Verbesserung: $\overline{r} = 17.89$,

2011) ergibt mit $p = 0.051$ keine Signifikanz. Versteht man die einfachere Heuristik als eine Baseline, wäre ein einseitiger Test angemessen und es ergäbe sich Signifikanz.

$\overline{r}_{MZ} = 16$, und $\overline{r}_{DZ} = 18.29$. Wir bekommen insgesamt 25 Verbesserungen bei nur 8 Verschlechterungen, d.h. signifikant mehr Verbesserungen.[63] Siehe auch Abbildung 3.22

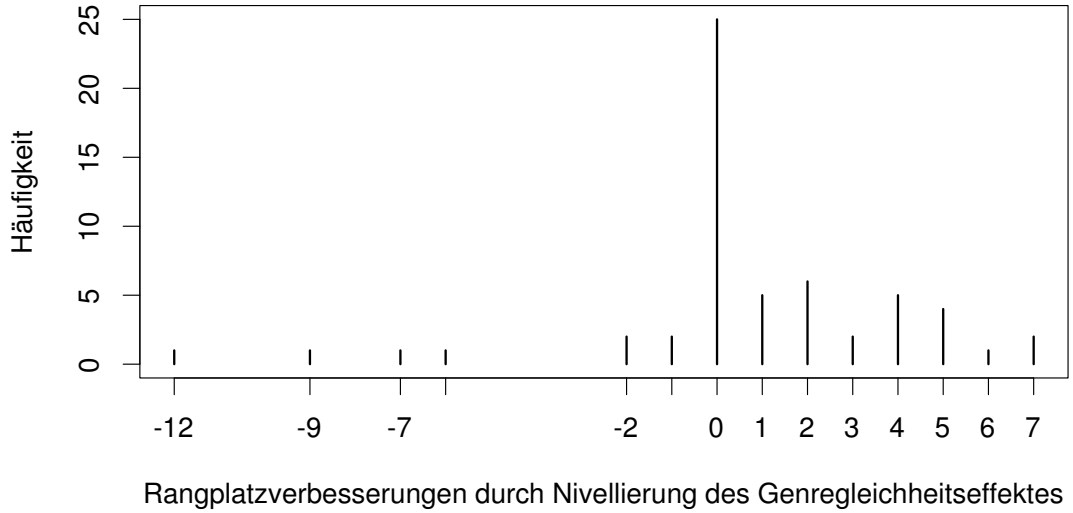


Abbildung 3.22: Rangplatzverschiebungen, die sich durch den einfachen Ausgleich des Genreeffektes ergeben. Die positiven Verschiebungen überwiegen signifikant.

Geht man noch einen Schritt weiter und bildet für jede Genre-Genre-Kombination eine eigene Gruppe, erhält man eine weitere Verbesserung: $\overline{r} = 16.26$, $\overline{r}_{MZ} = 10.6$, und $\overline{r}_{DZ} = 17.43$. Aber so stark diese Verbesserung aussieht, ist sie dennoch nicht nachweisbar signifikant.

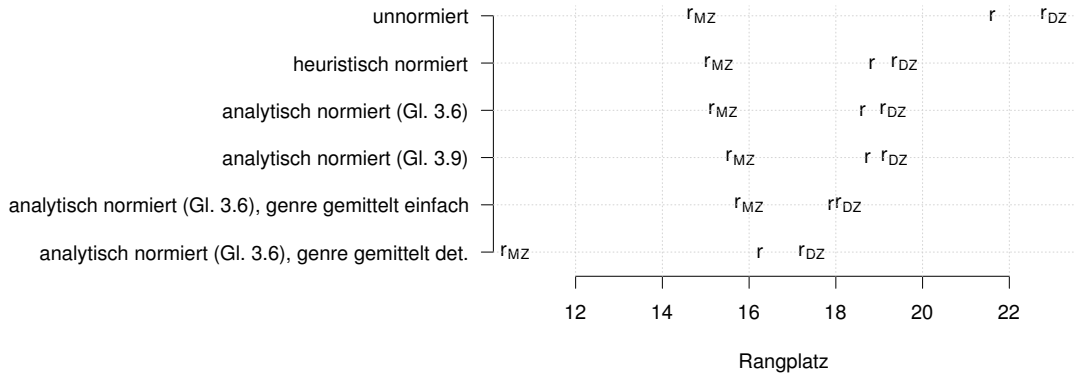
In Tabelle 3.8 ist die zuerst beschriebene, einfachere Variante der Genremittelung als „einfache Mittelung“ bezeichnet. Die Variante, die die jeweiligen Genre-Kombinationen berücksichtigt, trägt die Bezeichnung „detaillierte Mittelung“.

Ein solches Vorgehen, die Wechselwirkungen mit dem Genre der Texte zu quantifizieren und direkt herauszurechnen ist meines Wissens neu.

Nachdem nun einerseits das Korpus optimal gereinigt und andererseits die Ergebnisse in Bezug auf das stilometrische Signal gereinigt sind, kann nun die zentrale Frage gestellt werden, ob in S eine erbliche Komponente nachgewiesen werden kann.

Die oben zitierten Unterschiede zwischen $\overline{r}_{MZ}$ und $\overline{r}_{DZ}$ erwecken den Eindruck, kein Zufall zu sein. In allen von mir betrachteten Varianten des Korpus und Betrachtungsweise der Ergebnisse war die bessere Klassifzierbarkeit der eineiigen Zwillinge sichtbar. Für die optimale Aufbereitung der Daten zeigt Abbildung 3.23 den Effekt noch einmal graphisch. Alle betrachteten Repräsentationen hängen allerdings untereinander zusammen und taugen daher nicht als unabhängige Bestätigung. Um die Signifikanz des Befundes auf verlässliche Art zu überprüfen bin ich folgendermaßen vorgegangen:

[63] Binomialtest: $p = 0.005$



Normierung	Genre	$\overline{r}$	$\overline{r}_{DZ}$	$\overline{r}_{MZ}$
unnormiert	–	21.62	23.02	14.90
heuristisch normiert	–	18.84	19.58	15.30
analytisch (Gleichung 3.4)	–	18.63	19.32	15.40
analytisch (Gleichung 3.7)	–	18.74	19.35	15.80
analytisch (Gleichung 3.4)	gemittelt (einfach)	17.89	18.29	16.00
analytisch (Gleichung 3.4)	gemittelt (detailliert)	16.26	17.43	10.60

Tabelle 3.8: Rangplätze im Überblick. Das Bild visualisiert die Tabelle.

Zuerst gilt es die Tatsache in Rechnung zu stellen, dass alle Zwillingspaare im selben Jahr also auch zum selben Thema geschrieben haben. Man kann nun aber davon ausgehen, dass Aufsätze zum selben Stimulusmaterial tendenziell ein höheres S erreichen. Damit haben Zwillinge von vornherein eine höhere Wahrscheinlichkeit als zueinander gehörig klassifiziert zu werden als man aufgrund der Zahl der Dateien alleine schätzen würde. Dem kann man Rechnung tragen, indem von vornherein nur $S(T_1, T_2)$-Werte betrachtet werden, für die T_1 und T_2 das Thema teilen.

Aufgrund der eingeschränkten Möglichkeiten bekommt man nun $\overline{r} = 10.17$, $\overline{r}_{MZ} = 7.20$ und $\overline{r}_{DZ} = 10.79$. Auch unter diesen eingeschränkten Bedingungen schneiden die eineiigen Zwillinge also besser ab. $\overline{r}_{MZ}$ beruht auf insgesamt 10 Rangplätzen für die zehn eineiigen Zwillinge. $\overline{r}_{DZ}$ beruht auf 48 Rangplätzen: 42 für die 21 zweieiigen Zwillingspaare und 6 für das Triplett. beide Rangplatzlisten werden vereinigt und die Gesamtliste wieder willkürlich auf zwei Listen mit 10 und 48 Mitgliedern verteilt. Anschließend werden wiederum die mittleren Rangplätze $\overline{r}_{MZ}$ und $\overline{r}_{DZ}$ berechnet und ihre Differenz mit der für die echten Paarungen gemessenen Differenz von $10.71 - 7.20 = 3.51$ verglichen. Nach 10^4 Wiederholungen wird die Zahl der Fälle gezählt, in denen eine mindestens so große Differenz auftritt. Aus dem Anteil dieser Fälle lässt sich ein p-Wert schätzen. Es ergibt sich $p \approx 0.11$, und damit kein signifikanter Effekt. Es ist also nicht auszuschließen, dass der gemessene Unterschied zwischen ein- und zweieiigen Zwillingen zufällig zustande gekommen ist. Dennoch stützen die Ergebnisse die Hypothese, dass

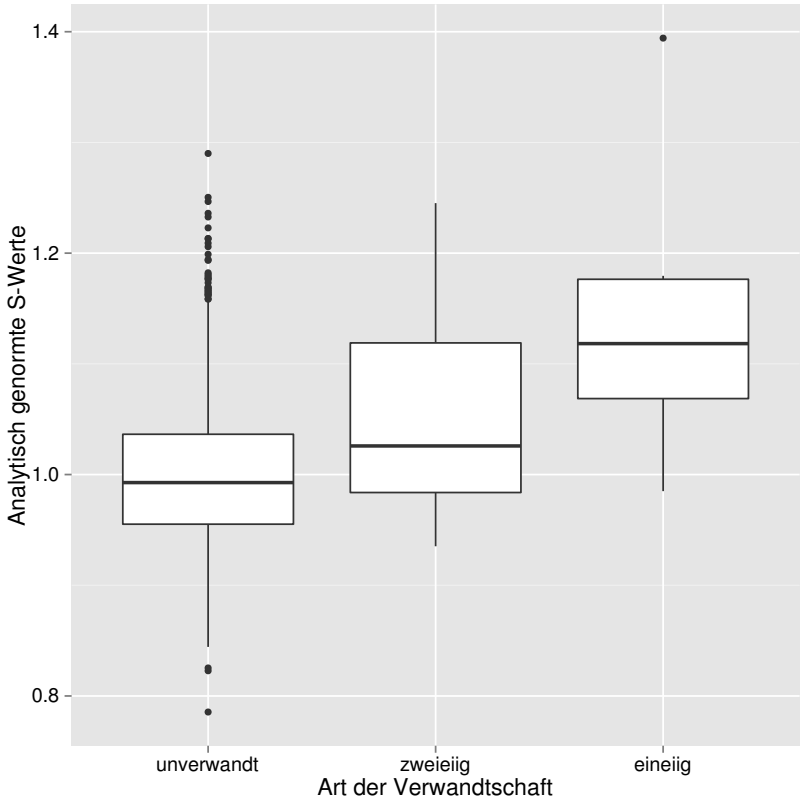


Abbildung 3.23: Die analytisch genormten S-Werte nach dem detaillierten herausmitteln der Genrebeziehungen in Abhängigkeit vom Verwandtschaftsgrad.

eineiige Zwillinge stilistisch ähnlicher schreiben als zweieiige Zwillinge eher, als dass sie sie widerlegen. Für eine Verifizierung sind allerdings mehr und vielleicht nicht ganz so komplexe Daten notwendig.

3.7 Zusammenfassung und Diskussion

In diesem Kapitel wird ein stilometrisches Verfahren auf Grundlage eines Textähnlichkeitsmaßes (S) entwickelt und evaluiert. Die herangezogenen Daten – *vollständige Substringhäufigkeiten* von Texten – wurden in ihrer Gesamtheit noch nicht für stilometrische Zwecke eingesetzt.

Einerseits wird gezeigt, dass der Algorithmus als *stilometrisches* Klassifikationsverfahren mit bisher veröffentlichten Verfahren, die meist auf tiefer annotierten Daten und unter Einsatz moderner Maschinenlernalgorithmen arbeiten, gut konkurrieren kann. Verfahren gegenüber, die ebenfalls nur auf Zeichenkettenhäufigkeiten zurückgreifen, erweist es sich als überlegen. Vor dem Hintergrund der Schlussfolgerung, die Juola (2006a) aus seinem *Ad-hoc Authorship Attribution competition* zieht, ist das ein nicht unerheblicher

Fortschritt:

> Unfortunately, another apparent result is that the high-performing algorithms appear to be mathematically and statistically (although not necessarily linguistically) sophisticated and to demand large numbers of features.

Im Gegensatz dazu ist die in der vorliegenden Arbeit vorgestellte Methode konzeptuell einfach: Die Häufigkeiten aller in beiden Texten vorkommenden Substrings werden verknüpft und aufsummiert. Dabei gibt es keinen freien Parameter und es wird kein maschinelles Lernverfahren eingesetzt.

Andererseits erlauben die Ergebnisse Rückschlüsse darauf, auf welcher sprachlichen Ebene sich stilometrisch relevante Informationen finden lassen und in welcher Form – linear oder logarithmisch – die in den Substringfrequenzen steckende Information am besten nutzbar wird.

Die bisher veröffentlichten Arbeiten zur Stilometrie lassen sich sehr grob in zwei Gruppen einteilen. In der einen Gruppe wird letztlich auf die eine oder andere Art ein Index für die Ähnlichkeit von Texten berechnet. Aufgrund dieses Indexes werden die Texte klassifiziert. Die andere Gruppe ordnet die Texte typischerweise in einem vieldimensionalen Raum an und verwendet ein maschinelles Lernverfahren für die Klassifizierung dieser Punkte im Raum. Der hier vorgestellte Algorithmus gehört zur ersten Gruppe.[64]

Das eingesetzte Ähnlichkeitsmaß hängt in seiner rohen Form unter anderem von der Länge der beiden verglichenen Texte ab. Da einerseits diese Abhängigkeit nicht parametrisiert werden kann und sie andererseits nicht der einzige einzeltextspezifische Einfluss auf das Ähnlichkeitsmaß ist, bedarf es einer heuristischen Normierung. Erst das normierte S lässt sich als ein Maß für die Ähnlichkeit der beiden Texte interpretieren, ohne Beiträge, die nur von einem der Texte bestimmt werden.

Insgesamt liegt die Performanz des Verfahrens gleich auf mit Verfahren, die zum einen etablierte und mächtige maschinelle Lernverfahren einsetzen und andererseits über den reinen Text hinaus auf weitere Informationen zugreifen, die teilweise erhebliches sprachliches Wissen implizieren, zum Beispiel Fehler im Text. Maschinelle Lernverfahren wie SVM nutzen die Beziehungen aller Texte eines Korpus zugleich aus, um Testtexte zu klassifizieren. Dies unterbleibt dagegen in vielen Algorithmen, die nicht auf ein maschinelles Lernverfahren setzen. Die im hier vorgestellten Ansatz verwendete Normierung besteht in ihrer Grundform in einer heuristisch begründeten einfachen Mittelung, die alle Texte mit einbezieht. Dadurch werden alle verglichenen Texte miteinander in Beziehung gesetzt. Die eigentliche Klassifizierung findet erst im Anschluss an die Normierung durch einen einfachen Vergleich der normierten Zahlen statt. Es lässt sich die Hypothese aufstellen, dass die Methode (unter anderem) durch dieses Charakteristikum so erfolgreich ist.

Das untersuchte Ähnlichkeitsmaß S liegt in 6 Varianten vor. Diese unterscheiden sich im Wesentlichen darin, ob die aufsummierten Frequenzdaten eines oder beider verglichenen Texte logarithmisch oder als absolute Zählungen in die Berechnungen eingehen.

[64] Im Rahmen der Untersuchung der *Federalist Papers* (Abschnitt 3.6.3) wird zwar am Rande mit einem maschinellen Lernverfahren experimentiert, dies ist aber kein Bestandteil des Verfahrens selber und variiert nur temporär und versuchsweise den eingesetzten Klassifikationsmechanismus.

Weitere Unterschiede betreffen nur die genaue Art der Aufsummierung. Die Varianten ermöglichen einen Vergleich der linearen und logarithmischen Form der Frequenzdaten in Bezug auf Trainings- und Testtexte.

Performanz und Verhalten des Verfahrens wurden auf fünf unterschiedlichen Korpora[65] und an vier unterschiedlichen Fragestellungen untersucht.

Juola (2004) ist ein speziell zu stilometrischen Testzwecken zusammengestelltes Korpus mit mehreren Subkorpora. Anhand dieses Korpus erweist sich die Methode zum einen als praktikabel. Darüber hinaus ergibt ein Vergleich mit Teilnehmern eines Wettbewerbs auf Grundlage dieses Korpus hervorragende Ergebnisse. Ein Vergleich der 6 definierten Varianten des verwendeten Ähnlichkeitsmaßes ergibt eine klare Überlegenheit für S_{log}, eine Variante, in der alle Substringfrequenzen logarithmisch eingehen.

Das Korpus von Baroni und Bernardini (2006) wurde in Bezug auf die Fragestellung erhoben, ob und bis zu welchem Grad es möglich ist, zwischen ursprünglich auf Italienisch verfassten Texten und Übersetzungen ins Italienische zu unterscheiden. Ich repliziere die Untersuchung von Baroni und Bernardini (2006). Auch in diesem Fall erweist sich das vorgestellte Verfahren als wettbewerbsfähig. Die Überlegenheit der logarithmischen Versionen von S wird bestätigt. Auch insgesamt ergibt sich wieder die bereits anhand von Juola (2004) bestimmte Performanzreihenfolge der Varianten. Das Korpus liegt nicht nur als Originaltext vor, sondern auch in diversen oberflächenannotierten Versionen. Es zeigt sich, dass die Methode in allen Textversionen eine beinahe identische Performanz besitzt (mit Ausnahme der `tag`-Version, die nur aus der Aneinanderreihung der POS-tags der inhaltstragenden Wörter besteht). Das ist insofern überraschend, als sehr viele stilometrische Verfahren auf der Annahme aufbauen, dass Funktionswörter die mächtigste Quelle stilistischer Information sind. Die Details der Auswertung legen nahe, dass für verschiedene Texte die für die Klassifikation nutzbare Information auf unterschiedlichen Ebenen angesiedelt sind. Außerdem zeigt sich, dass die verschiedenen Textrepräsentationen nur in der logarithmischen Darstellung der Daten gleich gut klassifiziert werden können. In linearer Darstellung erlauben die Funktionswörter in der Tat die effektivste Klassifizierung. Da der Logarithmus den selteneren Zeichenketten gegenüber den häufigen mehr Gewicht zubilligt, deutet die hohe Klassifikationsqualität für die übrigen Repräsentationen in logarithmischer Repräsentation darauf hin, dass in den selteneren n-Grammen erhebliche Information steckt, die leicht von den höherfrequenten n-Grammen verdeckt wird. In eine ähnliche Richtung weist die Beobachtung, dass in allen Repräsentationen die Performanz der Methode nachlässt, wenn die Token nicht in Originalreihenfolge erscheinen, sondern randomisiert sind. Durch die Randomisierung geht die Kohärenz der längeren Ketten verloren. Der Abfall der Klassifikationsqualität in diesem Fall ist ein Anzeichen für den Informationsgehalt eben dieser längeren n-Gramme.

Das ICLE-Korpus besteht aus Texten fortgeschrittener Englisch-Lerner. Das vorgestellte stilometrische Verfahren wird eingesetzt, um die Autoren nach ihrer Muttersprache zu klassifizieren. Auch hier besteht Vergleichbarkeit mit veröffentlichen Performanzwerten anderer Verfahren. Es ergibt sich, dass die S-basierte Klassifikation an-

[65] Juolas Testkorpus (Juola, 2004), das *Limes*-Korpus (Baroni und Bernardini, 2006), ICLE (Granger, 2003), die *Federalist Papers* (Hamilton et al., 2004) und das Zwillingskorpus von Mollet et al. (2010).

deren Methoden, die nur Substringfrequenzen verwenden, deutlich überlegen ist, selbst wenn diese maschinelle Lernverfahren einsetzen. Da die vorliegende Arbeit die erste ist, die die *vollständigen Substringhäufigkeiten* verwendet, während die konkurrierenden Ansätze jeweils nur einen kleinen Tei der häufigen und kurzen Zeichenketten verwerten, ist dies ein deutlicher Hinweis auf den nicht zu vernachlässigenden Informationsgehalt der selteneren und längeren Zeichenketten. Die Methode ist auch gegenüber Verfahren wettbewerbsfähig, die weitergehende annotierte Informationen hinzuziehen. Die hohe Sensibilität des Verfahrens zeigt sich darin, dass Details und Probleme des Korpus erkannt werden können, die in anderen Arbeiten unentdeckt bleiben.

Mit den *Federalist Papers* wird ein traditionelles Standardproblem der Stilometrie untersucht. Eine Besonderheit dieses Korpus ist seine extreme thematische und genrebezogene Konstanz. In Bezug auf die verwendete Methode ergibt sich ein besonderes Problem. In den bisherigen Untersuchungen war per Design sichergestellt, dass die Menge der Testtexte ausgewogen über die jeweiligen Kategorien verteilt war. So bildeten im Rahmen der *Translationese*-Untersuchung jeweils 15 Originale und 15 Übersetzungen die Testmenge. Hier nun ist es möglich und sogar sehr wahrscheinlich, dass alle umstrittenen Texte vom selben Autor stammen. In einer solchen Anordnung versagt die einfache Form der Normierung. Sie wird durch eine Normierung mit neutralen Hilfsdateien ersetzt. Hierfür kommen Texte aus dem BNC (Burnard, Lou , Hg.) zum Einsatz, also aus einer anderen Zeitstufe und Varietät. Die Qualität der so erzielten Ergebnisse ist hoch. Zum einen können die Resultate der bisherigen Forschung repliziert werden. Darüber hinaus fallen Besonderheiten eines der fünf Texte des dritten Autors James Jay auf, die in der stilometrischen Literatur sonst oft unerwähnt bleiben. Auch zwei der gesichert Hamilton bzw. Madison zugeordneten Texte verhalten sich abweichend. Diese Abweichung rührt möglicherweise aus thematischen Beziehungen unter den Texten her. Angesichts der extremen thematischen Gleichmäßigkeit des Korpus wäre eine Auflösung, die hoch genug ist, solch subtile Strukturen zu entdecken, bemerkenswert.

Das fünfte untersuchte Korpus (Mollet et al., 2010) dagegen stellt durch seine extreme thematische und genrebezogene *Heterogenität* und die Vielzahl der beitragenden Autoren einen Gegensatz zu den *Federalist Papers* dar. Die 55 Texte umfassende Sammlung wurde von 26 ein- und zweieiigen Zwillings- und Drillingspaaren verfasst. Es wird untersucht, ob die Texte der eineiigen Zwillinge ähnlicher sind als die der zweieiigen. Dies wäre ein starkes Indiz für die Erblichkeit von Stil, wie ihn das Ähnlichkeitsmaß S erfasst. Diese quantitativ vergleichende Fragestellung ist meines Wissens eine Neuheit in der Stilometrie. Auch wenn sich keine statistische Signifikanz nachweisen lässt, stützen die Daten die Erblichkeitshypothese sehr viel eher, als dass sie sie widerlegen. Die statistische Power[66] leidet unter der starken Heterogenität der Texte. Diese erlaubt es auf der anderen Seite, den Einfluss von Text*topic* und -*genre* zu untersuchen. Es zeigen sich stark gegenläufige Tendenzen. So reduziert eine Transformation der Texte in POS-Sequenzen zwar den Einfluss des *Topics* erheblich, verstärkt allerdings die *Genre*abhängigkeit der Textähnlichkeit. Derartige Effekte werden in der Stilometrie gewöhnlich nicht thema-

[66] Definiert als die Wahrscheinlichkeit, dass ein tatsächlich existierender Effekt sich auch in einem signifikanten Ergebnis äußert. Diese Wahrscheinlichkeit sinkt, wenn die Varianz in den Daten steigt.

tisiert.

Das Korpus eignet sich auch für eine Untersuchung eines allgemeinen Problems elizitierter Korpora: Formulierungen aus der Themenstellung werden als Ganzes in die produzierten Texte übernommen und haben das Potential, Forschungsergebnisse in schwer einschätzbarer Weise zu beeinflussen. In Bezug auf dieses Problem wird eine Filtermethode vorgestellt, die direkt auf den untersuchten vollständigen Substringhäufigkeiten aufbaut. Ihre Wirksamkeit wird empirisch nachgewiesen. Nebenbei ergibt sich eine operationalisierende Definition *unnatürlicher Wiederholungen* in einem Text, die eine gewisse allgemeine Gültigkeit haben könnte und die sich möglicherweise in vielfältigen Kontexten als nützlich erweist.

Darüber hinaus wird die bisher verwendete rein heuristische Normierung mit einem genauer begründeten Verfahren verglichen. Die Experimente zeigen eine leichte Überlegenheit der komplexeren Variante.

Es gibt eine sehr auffällige Parallelität der hier beschriebenen Ergebnisse zur *Stilometrie* mit den Schlüssen, die in Kapitel 2 aus den Analysen zur *Morphologischen Induktion* gezogen wurden. In beiden Fällen wirkt sich die logarithmische Transformation der Substringhäufigkeiten klar und konsistent positiv auf die Performanz des Algorithmus aus. Nun gibt die Logarithmierung den längeren und selteneren Substrings mehr Gewicht gegenüber den kürzeren und häufigeren. Daher weist die Überlegenheit der logarithmischen Repräsentation darauf hin, dass die selteneren Substrings mehr relevante Information enthalten als gemeinhin angenommen. Die traditionelle Unterrepräsentation der niedrigfrequenten Ereignisse[67] hat verschiedene Gründe:

Zum einen wurden diese Daten, teilweise sicherlich wegen ihrer extremen Masse, schlicht nicht betrachtet. Sie werden aber auch ausgeschlossen, da sie als vergleichsweise nutzlos gelten. Abbildung 3.12 und die anschließende Diskussion stützen die Hypothese, dass dies wiederum daran liegen könnte, dass diese Annahme für die untransformierten Häufigkeitsdaten tatsächlich zutrifft. Die Werthaltigkeit der vernachlässigten niedrigfrequenten Daten wird erst in logarithmischer Transformation deutlich.

Darüber hinaus ist gerade die Seltenheit der längeren Ketten in vielen Kontexten ein Problem, da sie zu *Data Sparseness* führt. Dies ist vor allem dann der Fall, wenn relative Häufigkeiten als Wahrscheinlichkeiten interpretiert werden, da dies schnell zur Notwendigkeit von Smoothingparametern führt. Für den hier vorgestellten Algorithmus dagegen tritt dieses Problem nicht auf, da von vornherein Substrings ausgeschlossen werden, die in einem der beiden Texte nicht vorkommen. Da alle übrigen Zeichenketten ganz unabhängig von ihrer Länge und Häufigkeit gleichermaßen eingehen, ist *Data Sparseness* insgesamt kein Problem.

Ein dritter Grund, warum seltene Strings häufig vernachlässigt werden, ist die Annahme, dass sich durch den Ausschluss längerer Ketten oder inhaltstragender Wörter der störende Einfluss des *Topics* unterdrücken lässt. Dies mag bis zu einem gewissen Grad zutreffen, zwei Einwände sprechen aber dennoch gegen ein solches Vorgheen. Einerseits

[67] Mit lediglich einer mir bekannten bemerkenswerten Ausnahme: „[...] both the high-frequency head of the rewrite frequency distribution s well as its low-frequency tail provide independent converging evidence for authorship [...]“ (Baayen et al., 1996). Vor allem die verwendeten (syntaktisch annotierten) Ausgangsdaten unterscheiden diese Arbeit allerdings stark von der vorliegenden.

verliert man durch diese Einschränkung große Mengen potentieller Information, während anderseits ein eventueller störender Einfluss des Text-*Genres* eher verstärkt wird. Die Untersuchungen zum Zwillingskorpus in Abschnitt 3.6.4 zeigen, dass es Möglichkeiten gibt, mit Störvariablen umzugehen, ohne, dass man von vorneherein potentiell informative Daten ausschließen muss.

4 Zusammenfassung und Ausblick

4.1 Zusammenfassung

In dieser Arbeit werden die Eigenschaften vollständiger Häufigkeitszählungen aller in einem Text enthaltenen Zeichenketten untersucht. Es kann gezeigt werden, dass diese Daten sowohl vom Anwendungsgesichtspunkt aus ein mächtiges Werkzeug darstellen, als auch theoretisch relevante Fragen aufzuwerfen und potenziell auch zu beantworten in der Lage sind.

Mit der *Morphologischen Induktion* (Kapitel 2) und der *Stilometrie* (Kapitel 3) werden zwei computer- und korpuslinguistisch relevante Fragestellungen untersucht.

Die *Morphologische Induktion* setzt es sich zum Ziel, rohen, unsegementierten Text mithilfe unüberwachter Lernverfahren in morphologische Einheiten zu zerlegen. Ein Beispiel wäre die automatisierte Deduktion der Tatsache, dass die Phrase `he|has|accomplish|ed` in die gezeigten Einheiten segmentierbar ist. In dieser Arbeit gelingt es, für diese Fragestellung einen stabilen und neuartigen Algorithmus vorzustellen. Seine Grundüberlegung besteht in der möglichst konsequenten Umsetzung des alten Gedankens, dass an den Grenzen *sprachlicher Segmente* die Vorhersagbarkeit der angrenzenden Zeichen abfällt (Harris, 1955). Darüber hinausgehendes sprachliches Wissen wird nicht implementiert. Verbleibende Mehrdeutigkeiten werden durch ein Rankingverfahren überwunden. Verschiedene Varianten dieses Rankingverfahrens werden sorgfältig miteinander verglichen. Sie entstehen aus der Variation von vier kategorialen Parametern.

Auf diesem Forschungsgebiet kann es aufgrund der Vielfalt der Fragestellungen, Zielsetzungen, Evaluationsmethoden und Datensätzen aktuell keine verlässliche quantitative Vergleichbarkeit geben. Sehr ungefähr liegen die ermittelten Performanzwerte im oberen Bereich der veröffentlichten Zahlen. In Deutsch und Englisch liegt der Anteil der als Segmentgrenzen erkannten Leerzeichen bei 95-96%, für das Türkische leicht darunter.

Die meisten der bisher veröffentlichten Arbeiten beschränken sich auf die ausschließliche Segmentierung komplexer Wörter. Die Zielsetzung meines Verfahrens dagegen beinhaltet nicht nur eine Segmentierung von Texten in *minimale sprachliche Segmente* (≈Morpheme), sondern darüber hinaus ihre Zusammenordnung zu Einheiten höherer Ordnung. So soll der Algorithmus erkennen, dass die beiden Zeichenketten `accomplish` und `ed` zusammen das Wort `acomplished` bilden.

Ebenfalls im Kontrast zum Großteil der Arbeiten ist kein tokenisierter Text oder eine Wortliste der Input des Verfahrens, sondern längere Textabschnitte, gegebenenfalls Sätze. Dies macht einerseits jedes Preprocessing (über diese rudimentäre Zerlegung hinaus) unnötig, andererseits ist das Verfahren damit auch ohne Anpassungen für Schrift-

systeme geeignet, die keine Leerzeichen kennen. Ein weiteres Alleinstellungsmerkmal ist die Art und Weise und die Radikalität, mit der der Kontext mit in die Segmentierungsentscheidungen des Systems eingeht. Einerseits wird keine obere Grenze für die Länge der berücksichtigten Zeichenketten gesetzt. Stattdessen werden Zeichenketten beliebiger Länge als Informationsquelle herangezogen. Die verwendete Kontextinformation geht aber über die Länge der längsten sich wiederholenden Zeichenkette noch hinaus. Bei der Auswahl der besten Segmentierung werden Zusammenhänge über die gesamte Länge des zu segmentierenden Satzes oder Absatzes mit einbezogen. So spielt auch das Zusammenspiel verschiedener Teile eines Satzes eine Rolle für die letztendliche Segmentierung.

Nicht nur diese Eigenschaften machen den vorgestellten Algorithmus zu einem Kandidaten für eine mächtige und vollständig sprachunabhängige Segmentierungsmethode. Bei der Auswahl der optimalen Segmentierung spielen insgesamt 4 kategoriale Parameter eine Rolle. P_L steuert die Bewertung der Einzelsegmente selbst (`accomplish` bzw. `ed`). P_T betrifft den Beitrag, die Teilsegmente zur Beurteilung eines übergeordneten Segmentes leisten. Im Beispiel könnte das die Frage betreffen, was `accomplish` und `ed` zum Ranking von `accomplished` beitragen. P_F wiederum ist für die Einschätzung von Folgen von Segmente verantwortlich, z.B. der Segmentfolge `he|has|accomplished` gegenüber einer denkbaren Alternative `heh|as|accomplished`. P_4 spielt nur dann eine Rolle, wenn sich ein übergeordnetes Element wie `accomplished` aus mehreren Segmenten bilden lässt, zum Beispiel aus `accomplish` und `ed` oder aus `accomp` und `lished`. Nun werden diese vier kategorialen Parameter $P_{L,T,F,4}$ zwar miteinander zu einer Vielzahl von Verfahrensvarianten kombiniert. Auf den untersuchten Korpora erweist sich jeweils derselbe Parametersatz als optimal. Legt man sich auf diesen Parametersatz fest, so bleibt keine weitere Sprachabhängigkeit mehr und jeder Text kann direkt segmentiert werden. Diese Sprachunabhängigkeit der optimalen Parameter ist für sich selbst genommen wiederum ein interessantes Ergebnis, da sie auf ähnliche Strukturen in typologisch sehr unterschiedlichen Sprachen hindeuten könnte.

Evaluiert wird der Algorithmus nicht nur an den drei Sprachen Deutsch, Englisch und Türkisch, sondern auch mit drei unterschiedlichen Evaluationsverfahren. Ein Problem stellt das Fehlen eines vertrauenswürdigen Goldstandards dar. Dies hat zwei wesentliche Gründe: Einerseits ist die Definition der morphologischen Grundeinheiten stark theorieabhängig, was die Zerlegung eines konkreten Textes und den Begriff des *Goldstandards* an sich fragwürdig erscheinen lässt. Aber auch wenn man sich auf eine bestimmte Auffassung von Morphologie festlegt, ist es immer noch ein sehr schwieriges und aufwendiges Unterfangen, größere Evaluationskorpora in mehreren Sprachen zu erstellen. Daher gehe ich in Abschnitt 2.6.2 zuerst einen anderen Weg. Es wird ausgenutzt, dass in aller Regel jedes Leerzeichen im Text einer Segmentgrenze entspricht. Da der Algorithmus selbst über dieses Wissen nicht verfügt, ergibt sich unmittelbar eine große Untermenge auswertbarer Segmentgrenzen. Um darüber hinaus wenigstens einen Einblick in die Performanz des Algorithmus in Bezug auf wortinterne Segmentgrenzen zu gewinnen, wird in Abschnitt 2.6.3 für das deutsche Subkorpus ein kleiner Goldstandard erstellt und ausgewertet. Die unvermeidliche Theorieabhängigkeit wird durch die unabhängige Befragung dreier Experten quantifizierbar gemacht. Es ergibt sich mit etwa 15% ein erstaunlich

hoher Anteil an divergierenden Einzelentscheidungen der Experten. Zur Evaluation werden die üblichen Performanzmaße *Recall*, *Precision* und f-Measure angepasst, um dieser Variabilität Rechnung tragen zu können.

Für diese beiden quantitativen Analysen werden sowohl *lineare gemischte Modelle* als auch *generalisierte gemischte Modelle* eingesetzt. *Lineare gemischte Modelle* bieten eine flexible Beschreibung normalverteilter Daten in Abhängigkeit von verschiedenen Klassen von Variablen wie der Satzlänge auf der einen Seite und zufälligen Einflussgrößen wie der spezifischen aber unvorhersagbaren Schwierigkeit eines bestimmten Satzes auf der anderen. *Generalisierte Modelle* sind darüber hinaus in der Lage, nicht nur normalverteilte Daten zu beschreiben, sondern auch die Wahrscheinlichkeit von Ereignissen zu modellieren. Ein Beispiel hierfür ist das Auftreten eines Klassifikationsfehlers an einer bestimmten Stelle im zu segmentierenden Text. Die erstmalige Anwendung dieser andernorts etablierten Methoden auf dem Gebiet der *Morphologischen Induktion* macht es möglich, aus dem Vergleich der vielen untersuchten Parameterkonstellationen theoretisch relevante Ergebnisse abzuleiten.

So erweisen sich die Ergebnisse als rechts-links-asymmetrisch. Das heißt, der linke und der rechte Rand einer Zeichenkette haben nicht dieselbe Mächtigkeit für die Entscheidung, ob es sich dabei um ein *sprachliches Segment* handelt oder nicht. Dieses Phänomen taucht an zwei unterschiedlichen Stellen auf. Einerseits beeinflusst es den Parameter P_L, der den einzelnen Zeichenketten einen Güteindex zuweist. Andererseits hat es eine Wirkung auf den Parameter P_4, der festlegt, welche Kindsegmente gewählt werden, falls es hier mehrere Möglichkeiten gibt. Bemerkenswerterweise geht die Asymmetrie in beiden Fällen in eine unterschiedliche Richtung. Konkret bedeutet dies, dass für die korrekte Beurteilung einzelner Segmente die Frequenzverhältnisse am hinteren Ende entscheidender sind als am vorderen Ende. Umgekehrt ist es bei der Auswahl unter verschiedenen Möglichkeiten der hierarchischen Aufspaltung vorteilhafter, die Frequenzen an den vorderen Segmentgrenzen auszuwerten.

Derartige Asymmetrien sind zwar zu erwarten, da Sprache immer eine eindeutige zeitliche Achse hat. Dennoch ist dies meines Wissens die erste Untersuchung, in der ein solcher Effekt in den Daten aufscheint. Dies unterstreicht die feine Auflösung der verwendeten Modelle. Es bleibt als Aufgabe für zukünftige morphologische Arbeiten, die Details derartiger Asymmetrien zu beschreiben, zu modellieren und zu erklären.

Zwei der untersuchten Parameter, P_L und P_T, erweisen sich in ihrem Verhalten als relativ stabil von Sprache zu Sprache, im Gegensatz zu P_F.[1] Während P_L und P_T die verschiedenen Segmentierungen lokal begrenzt bewerten, ist P_F für die globale Kombination der lokalen Segmente zu einer Segmentierung des gesamten Satzes verantwortlich. Ob sich diese unterschiedliche Variabilität in den lokalen und globalen Verhältnissen in einem breiteren Sprachenspektrum bestätigen lässt, ist eine hoch interessante Frage.

In Abschnitt 2.6.4 werden diese Untersuchungen durch eine manuelle Inspektion eines Querschnitts der entstehenden Segmente ergänzt. Hier liegt ein Schwerpunkt auf der Analyse der vorkommenden Fehler. Es erweist sich zum einen, dass ein Großteil der Fehler

[1] Für P_4 wurde kein Sprachvergleich durchgeführt. Dieser Parameter hat einen so geringen Einfluss, dass er nur anhand des kleinen deutschen Goldstandards untersucht wird.

lediglich in einer Übersegmentierung besteht, nicht in einer eigentlichen Fehlanalyse der Struktur. Zum anderen erkennt man hier, dass ein Verfahren, das auf der untersten Ebene der *minimalen sprachlichen Segmente* gut funktioniert, auf höheren Ebenen zu systematischen Problemen führt. Während *minimale sprachliche Segmente* gut erkannt werden, entspricht ein großer Teil der längeren Segmente eher einer Musterkonstruktion mit Leerstelle (`according to` X).

Die allgemeinste und vielleicht weitreichendste Schlussfolgerung lässt sich aus der Wirkung des Parameters P_L ableiten. P_L entscheidet darüber, in welcher Form die ursprünglichen Frequenzdaten der Substrings des Trainingstextes in die Entscheidung für die letztendliche Segmentierung eingehen. Hier gibt es drei Möglichkeiten: Die Frequenzinformation bleibt unberücksichtigt, sie geht linear ein, oder sie geht logarithmisch ein. Es ist ein sehr stabiles Ergebnis der Evaluation, dass die logarithmische Form den beiden anderen überlegen ist. Die logarithmische Transformation macht Messgrößen unterschiedlicher Größenordnungen miteinander vergleichbar. Dies bedeutet in Konsequenz, dass kleinere Zählungen gegenüber den größeren an Gewicht gewinnen. Daher kann man die Überlegenheit der logarithmierten Form der Daten so deuten, dass es sich bezahlt macht, die niedrigfrequenten Strings ebenfalls in die Analyse mit einzubeziehen.

Nach diesen in Kapitel 2 dargestellten Untersuchungen zur *morphologischen Induktion* wendet sich Kapitel 3 dem Forschungsgebiet der *Stilometrie* zu. Aus den Daten, die den Untersuchungsgegenstand dieser Arbeit darstellen – den *vollständigen Substringhäufigkeiten* von Texten – kann ein Textähnlichkeitsmaß gewonnen werden, das sich für effiziente Stilometrie eignet. Die Methode wird auf einer breiten Datenbasis aus fünf unterschiedlichen Korpora[2] evaluiert. Die dabei untersuchten Fragestellungen umfassen die Automatische Autorenbestimmung, die Klassifizierung in Übersetzungen und Originale, die Klassifizierung nach der Muttersprache des Autors und die Untersuchung des Stils von Zwillingspaaren. Die ersten drei Aufgaben ordnen sich in den üblichen Rahmen stilometrischer Klassifizierung ein, die letzte Aufgabe aber ist innerhalb der *Stilometrie* ein spezieller Fall. Der Nachweis der Vererbbarkeit von Stil erfordert nicht nur den qualitativen Nachweis der Ähnlichkeit der Texte von Zwillingen, sondern darüber hinaus den quantitativen Nachweis, dass eineiige Zwillinge ähnlicher schreiben als zweieiige. In dieser Arbeit wird meines Wissens eine derartige Fragestellung das erste Mal untersucht.

Wo ein Vergleich möglich ist, ergeben sich in allen beschriebenen Szenarien jeweils Performanzwerte, die konkurrierenden Ansätzen entweder überlegen waren (Abschnitt 3.5), oder ebenbürtig (Abschnitte 3.6.1 und 3.6.2). Dabei ist auffällig, dass in den verglichenen Arbeiten jeweils die Informationen mehrerer Annotationsebenen und/oder weiterer Informationsquellen verbunden und etablierte maschinelle Lernverfahren eingesetzt werden. Das von mir entwickelte Verfahren verzichtet auf beides.

Das definierte Textähnlichkeitsmaß existiert in 6 Varianten, die sich darin unterscheiden, ob die Substringhäufigkeiten linear oder logarithmisch eingehen und auf welche Art sie genau miteinander verbunden werden. Es ergibt sich auf allen Daten und unter allen Fragestellungen eine stabile Reihenfolge: Überlegen sind die Maße, in die die Frequenzen

[2] Juolas Testkorpus (Juola, 2004), das *Limes*-Korpus (Baroni und Bernardini, 2006), ICLE (Granger, 2003), die *Federalist Papers* (Hamilton et al., 2004) und das Zwillingskorpus von Mollet et al. (2010).

der verglichenen Texte logarithmisch eingehen.[3]

Dieses sehr klare empirische Ergebnis ist verblüffend ähnlich zu den Ergebnissen der *Morphologischen Induktion* aus Kapitel 2. Wieder kann der Schluss gezogen werden, dass es sich lohnt, die selteneren und längeren Substrings gegenüber den kurzen häufigen aufzuwerten. Flankiert werden die Schlussfolgerungen aus der Überlegenheit der logarithmischen Transformation von einer anderen Beobachtung. Im Rahmen der *Translationese*-Untersuchung (Abschnitt 3.6.1) werden verschiedene Repräsentationen desselben Korpus untersucht. Es ist ein bemerkenswertes Resultat, dass Repräsentationen, die nur aus den Oberflächenformen der Funktionswörter bestehen, ebenso gut abschneiden wie die, die sich aus der komplementären Menge der inhaltstragenden Wörter zusammensetzen. Funktionswörter sind tendenziell kurz und häufig, während die Mehrzahl der lexikalisch bedeutsamen Wörter länger und viel seltener ist. Sehr viele stilometrische Ansätze bedienen sich ausschließlich der Funktionswörter als alleiniger Datenbasis. Die hier vorgestellten Ergebnisse deuten darauf hin, dass diese Beschränkung nicht in jedem Fall angemessen ist. Es zeigt sich aber, dass in linearer Repräsentation genau die Funktionswörter die stilometrisch wirksamsten sind. Es ist plausibel, dass die Unterschätzung der langen Zeichenketten aus der überwiegenden Verwendung linearer Häufigkeiten in der Forschung herrührt.

In einem größeren Zusammenhang gesehen, bestätigen diese Ergebnisse die Vermutungen, die in der Einleitung (Kapitel 1) aus den Forschungen zur Korrelationsstruktur von Texten (Schenkel et al., 1993; Amit et al., 1994; Ebeling und Pöschel, 1994; Ebeling und Neiman, 1995; Ebeling et al., 1995; Montemurro und Pury, 2002; Altmann et al., 2012) abgeleitet wurden. Wie dort dargelegt, werden in den zitierten Arbeiten empirische Befunde präsentiert, die zeigen, dass es keine typische Skala gibt, auf der Korrelationen zwischen zwei Textstellen abklingen. Entsprechend sollte auch bei der Entwicklung sprachverarbeitender Algorithmen keine Skala eingeführt werden, zum Beispiel über eine feste Längen- oder Häufigkeitsschwelle, oberhalb oder unterhalb derer Zeichenketten von vorneherein als unwichtig betrachtet werden. Genau dies geschieht in vielen Arbeiten zu den hier behandelten Themenbereichen, wie das in der Einleitung zitierte Beispiel von Teahan (2000) zeigt. Stattdessen lässt sich aus verschiedenen Aspekten der sehr unterschiedlichen Untersuchungen dieser Arbeit eine fundamentale Schlussfolgerung ziehen: In den längeren und selteneren Zeichenketten steckt strukturelle Information, die in der Lage ist, die Performanz sprachverarbeitender bzw. -analysierender Algorithmen deutlich zu verbessern. Es entspricht der Grundüberzeugung hinter dieser Arbeit, dass die Rücksichtnahme auf empirische Erkenntnisse zur Struktur von Sprache anwendungsorientierten Algorithmen entscheidende Vorteile bringt und dass andersherum betrachtet aus der Performanz der Algorithmen linguistische Schlüsse gezogen werden können und sollen. Als entsprechend fruchtbar könnte es sich erweisen, einerseits den selteneren Ereignissen und ihren Wechselwirkungen in der linguistischen Forschung mehr Gewicht einzuräumen und andererseits die skalenfreie Natur der sprachinternen Dynamik von vorneherein in die Überlegungen und Modelle mit einzubeziehen. Altmann

[3] Am besten schneidet das Maß ab, das die Logarithmen der Produkte der beiden Frequenzen $\log\left(F_1(s) \cdot F_2(s) + 1\right)$ aufsummiert.

et al. (2012) fassen diese Forderung und den derzeit existierenden Widerspruch folgendermaßen zusammen:

> Understanding how language processes long-range correlations, an ubiquitous signature of complexity present in human activities (Voss und Clarke, 1975; Gilden et al., 1995; Yamasaki et al., 2005; Rybski et al., 2009; Kello et al., 2010) and in the natural world (Press, 1978; Kaneko und Li, 1992; Peng et al., 1992; Voss, 1992), is an important task towards comprehending how natural language works and evolves. This understanding is also crucial to improve the increasingly important applications of information theory and statistical natural language processing, which are mostly based on short-range-correlations methods (Manning und Schütze, 1999; Stamatatos, 2009; Oberlander und Brew, 2000; Usatenko und Yampol'skii, 2003).

Diese Arbeit soll einen Schritt in diese Richtung darstellen.

4.2 Ausblick

Aus den Ergebnissen dieser Arbeit erwachsen zahlreiche neue Fragen, die Beachtung verdienen.

Für den in Kapitel 2 dargestellten Algorithmus zur *Morphologischen Induktion* scheint es zum einen lohnend, die bestehenden Untersuchungen in Tiefe und Breite auszudehnen. Mit dem gewonnen Wissen wäre es denkbar, neue, von vornherein optimalere Strategien für die Parameter $P_{L,T,F \text{ und } 4}$ zu entwickeln. Dies wäre zum einen geeignet, das Verfahren zu optimieren und verspricht überdies weitere linguistisch relevante Erkenntnisse, zum Beispiel in Hinblick auf weitere Unterschiede und Gemeinsamkeiten des Verhaltens des Algorithmus in Bezug auf bisher nicht untersuchte Sprachen.

Vielversprechend scheint auch eine genauere Untersuchung der beobachteten *forward-backward*-Asymmetrien. Hier wäre es einerseits wichtig, mehr Material zu weiteren Sprachen zu erhalten, andererseits, nach weiteren Manifestationen derartiger Phänomene zu suchen. Es sollte sich für die Entwicklung der Theorie als fruchtbar erweisen, möglichst verlässliche empirische Daten zusammenzutragen, an denen sich neue Modelle entwickeln und testen lassen.

Eine weitere Informationsquelle zur optimalen Performanz des Algorithmus und den Details der Wechselwirkung der Parameter wäre die Erstellung eines Goldstandards, der eine vollständige Segmentierung von der Satzebene aus bis hinunter zu den *minimalen sprachlichen Segmenten* vorgibt. Die Erfahrungen mit dem in Abschnitt 2.6.3 untersuchten Goldstandard zeigen, dass sich auch aus einem extrem kleinen Vorrat positiver Beispiele sehr weitreichende Ergebnisse ableiten lassen. Dies lässt eine mehrsprachige und vollständige Evaluation nach diesem Schema realistisch erscheinen.

Ein Ansatzpunkt für eine wesentliche Weiterentwicklung des Algorithmus wäre seine Ergänzung um eine kategoriale Komponente (vgl. Abschnitt 2.6.4). Aus dem Verfahren in seiner jetzigen Form lässt sich die Schlussfolgerung ableiten, dass die Performanzgrenze für ein rein segmentierendes Vorgehen erreicht ist. Sowohl aus den Erfolgen als auch

aus den Grenzen des vorgestellten Algorithmus für *Morphologische Induktion* können Hinweise darauf abgeleitet werden, wie die nächste Entwicklungsstufe des Verfahrens aussehen könnte. Der vorgestellte Algorithmus funktioniert gut für die Identifizierung von Morphemen oder *minimalen sprachlichen Segmenten* in der von mir eingeführte Terminologie. Allerdings scheitert der Algorithmus gelegentlich daran, morphologisch oder syntaktisch unmögliche Abfolgen von Segmenten auszuschließen. Ein Beispiel hierfür bietet die Segmentierung `myhusband|is|ing|re|at|pain` (s. Seite 124). Auf höherer Ebene, beim Zusammenordnen der minimalen Segmente zu größeren Einheiten, ergeben sich strukturell andere Probleme. Betrachten wir als Beispiel eine häufige Konstruktion des Englischen: `according to` X, wobei X die Kategorie möglicher Folgeelemente repräsentiert, wie `you`, `plan` oder `Mr. Smith`. Da der Algorithmus kein Wissen über derartige Kategorien hat und diese auch nicht zu lernen im Stande ist, kann er ihre Elemente nur als unterschiedlich wahrnehmen. Infolgedessen neigt er dazu, eine die Zusammengehörigkeit von schablonenhaften Ausdrücken wie `according to` mit der Realisierung ihrer variablen Leerstelle X zu übersehen und beide Teile zu trennen. Das Zusammenordnen entstehender Segmente zu Kategorien und das Lernen von Regeln, nach denen diese Kategorien verbunden werden können, wäre daher möglicherweise ein Schritt in Richtung auf ein wesentlich mächtigeres Segmentierungsverfahren. Auf diese Weise könnten die beiden wesentlichen Fehlerquellen des aktuellen Algorithmus vermieden werden: Fehler auf unterster Ebene durch das völlige Fehlen einer morphosyntaktischen Komponente und Fehler auf höherer Ebene, die aus dem Fehlen eines kategorienbildenden Mechanismus herrühren.

Vom Anwendungsaspekt aus betrachtet, wäre auf kurze Sicht ein trivialeres Vorgehen möglicherweise vielversprechender: Viele Sprachen verwenden Leerzeichen, um Wörter zu trennen. Diese könnten dem Algorithmus zwingend als Segmentgrenzen vorgeschrieben werden. Es ist wahrscheinlich, dass dies zu einer starken Reduzierung der anfänglichen Mehrdeutigkeit der Segmentierungen und somit zu einer substanziellen Performanzverbesserung führt.

In Bezug auf die stilometrischen Untersuchungen (Kapitel 3) wäre ein naheliegender nächster Schritt die Bearbeitung der Frage, welche Zeichenketten genau bei den verschiedenen stilometrischen Aufgabestellungen zur erfolgreichen stilometrischen Klassifikation erforderlich und hilfreich sind. Wenn sich hier klare Strukturen erkennen lassen, wäre dies nicht nur für die Stilometrie selbst von Interesse, sondern könnte auch andere linguistische Teildisziplinen betreffen. So wäre es für die Lernersprachenforschung eine verwertbare Information, welche Zeichenketten auf welcher Ebene des Textes die Muttersprache eines Sprechers erkennbar machen. Ähnlich interessant wäre es für die Übersetzungsforschung, genauer zu erfahren, welche (Klassen von) Zeichenketten spezifisch für Originale bzw. Übersetzungen sind. Beachtung verdient hierbei insbesondere die weitere Untersuchung der Frage, ob sich die Hypothese bestätigt, dass die Hinweise in verschiedenen Texten auf verschiedenen Ebenen liegen wie dies die Ergebnisse in Abschnitt 3.6.1 vermuten lassen.

Literaturverzeichnis

Abbasi, Ahmet und Chen, Hsinchun: Writeprints: A stylometric approach to identity-level identification and similarity detection. In: *ACM Transactions on Information Systems*, Band 26(2), 2008.

Adair, Douglass: The authorship of the disputed federalist papers. In: *The William and Mary Quarterly*, Band 1(3):S. 97–122 und 235–264, 1944. 2 Teile.

Altmann, Eduardo G., Cristadoro, Giampaolo und Esposti, Mirko Degli: On the origin of long-range correlations in texts. In: *Proceedings of the National Academy of Sciences*, Band 109(29):S. 11582–11587, 2012.

Amit, M.[4], Shmerler, Y., Eisenberg, Eli, Abraham, M. und Shnerbh, Nadav: Language and codification dependence of long-range correlations in texts. In: *Fractals*, Band 2:S. 7–15, 1994.

Ando, Rie Kubota und Lee, Lillian: Mostly-unsupervised statistical segmentation of japanese kanji sequences. In: *Natural Language Engineering*, Band 9(2):S. 127–149, 2003.

Argamon, Shlomo, Akiva, Navot, Amir, Amihood und Kapah, Oren: Efficient unsupervised recursive word segmentation using minimumdescription length. In: *In Proc. 20th International Conference on Computational Linguistics g(Coling-04)*. 2004, S. 22–29.

Argamon, Shlomo, Koppel, Moshe und Shimoni, Anat Rachel: Gender, genre, and writing style in formal written texts. In: *Text*, Band 23(3):S. 321–346, 2003.

Baayen, R. Harald: *Word Frequency Distributions*. Kluwer, Dordrecht, 2001.

Baayen, R. Harald: *Analyzing Linguistic Data – A practical introduction to statistics*. Cambridge University Press, 2008.

Baayen, R. Harald, van Halteren, Hans und Tweedie, Fiona: Outside the cave of shadows: Using syntactic annotation to enhance authorship attribution. In: *Literary and Linguistic Computing*, Band 11(3):S. 121–131, 1996.

Baker, Mona: Corpus linguistics and translation studies – implications and applications. In: *Text and Technology: In Honour of John Sinclair*, John Benjamins, Amsterdam, S. 233–250. 1993.

[4]Nicht alle Vornamen waren ermittelbar.

Baker, Mona: Corpus-based translation studies: The challenges that lie ahead. In: Somers, Harold L. (Hg.) *Terminology, LSP and Translation: Studies in Language Engineering in Honour of Juan C. Sager*, John Benjamins, Amsterdam, S. 175–186. 1996.

Baroni, Marco: Distribution-driven morpheme discovery: A computational/experimental study. In: Booij, Geert und van Marle, Jaap (Hg.) *Yearbook of Morphology 2003*, Springer, Dordrecht, S. 213–248. 2003.

Baroni, Marco: Distributions in text. In: Lüdeling, Anke und Kytö, Merja (Hg.) *Corpus Linguistics. An International Handbook.*, Mouton de Gruyter, Berlin, Handbücher zur Sprach- und Kommunikationswissenschaft, S. 803–821. 2008.

Baroni, Marco und Bernardini, Silvia: A new approach to the study of translationese: Machine-learning the difference between original and translated text. In: *Literary and Linguistic Computing*, Band 21(3):S. 259–274, 2006.

Baroni, Marco, Matiasek, Johannes und Trost, Harald: Unsupervised learning of morphologically related words based on orthographic and semantic similarity. In: *ACL Workshop Morphol. & Phonol. Learning*. 2002, S. 48–57.

Bates, Douglas M., Maechler, Martin und Bolker, Ben: lme4: Linear mixed-effects models using S4 classes. URL `http://CRAN.R-project.org/package=lme4` (besucht am 12.10.2012), 2011. R package version 0.999375-39.

Bauer, Laurie: *Introducing Linguistic Morphology*. Edinburgh University Press, Edinburgh, zweite Auflage, 2003.

Bebel, August: Aus meinem Leben – Erster Teil. URL `http://www.gutenberg.org/files/12267/12267-8.txt` (besucht am 15.10.2012), 2004a.

Bebel, August: Aus meinem Leben – Zweiter Teil. URL `http://www.gutenberg.org/files/13690/13690-8.txt` (besucht am 15.10.2012), 2004b.

Benedetto, Dario, Caglioti, Emanuele und Loreto, Vittorio: Language trees and zipping. In: *Physical Review Letters*, Band 88(4):S. 048702, 2002a.

Benedetto, Dario, Caglioti, Emanuele und Loreto, Vittorio: On J. Goodman's comment to „Language Trees and Zipping". URL `http://www.citebase.org/abstract?id=oai:arXiv.org:cond-mat/0203275` (besucht am 12.10.2012), 2002b.

Benedetto, Dario, Caglioti, Emanuele und Loreto, Vittorio: Benedetto, Caglioti, and Loreto Reply:. In: *Phys. Rev. Lett.*, Band 90(8):S. 089804, 2003.

Bernardini, Silvia und Baroni, Marco: Spotting translationese: A corpus-driven approach using support vector machines. In: *Proceedings of Corpus Linguistics 2005*. 2006.

Best, Karl-Heinz: Zur Länge von Morphen in deutschen Texten. In: Best, Karl-Heinz (Hg.) *Häufigkeitsverteilungen in Texten*, Peust & Gutschmidt, Göttingen, S. 1–14. 2001.

Biber, Douglas: A register perspective on grammar and discourse: Variability in the form and use of english complement clauses. In: *Discourse Studies*, Band 1(2):S. 131–150, 1999.

Biber, Douglas und Barbieri, Federica: Lexical bundles in university spoken and written registers. In: *English for Specific Purposes*, Band 26:S. 263–286, 2007.

Borin, Lars und Prütz, Klas: Through a glass darkly: part of speech distribution in original and translated text. In: Daelemans, Walter, Sima'an, Khalil, Veenstra, Jorn und Zavrel, Jakub (Hg.) *CLIN*. Rodopi, 2000, Band 37 von *Language and Computers - Studies in Practical Linguistics*, S. 30–44.

Bortz, Jürgen: *Statistik für Sozialwissenschaftler*. Springer, Heidelberg, 6. Auflage, 2005.

Bosch, Robert A. und Smith, Jason A.: Separating hyperplanes and the authorship of the disputed federalist papers. In: *American Mathematical Monthly*, Band 105(7):S. 601–608, 1998.

Brent, Michael R.: Minimal generative models: A middle ground between neurons and triggers. In: *Proceedings of the 15th Annual Conference of the CognitiveScience Society*. Lawrence Erlbaum Associates, 1993, S. 28–36.

Brent, Michael R.: An efficient, probabilistically sound algorithm for segmentation and word discovery. In: *Machine Learning*, Band 34(1–3):S. 71–105, 1999.

Brent, Michael R., Murthy, Sreerama K. und Lundberg, Andrew: Discovering morphemic suffixes a case study in MDL induction. In: *Fifth International Workshop on AI and Statistics*. 1995, S. 264–271.

Burnard, Lou (Hg.): The British National Corpus Users Reference Guide. URL `http://www.natcorp.ox.ac.uk/docs/userManual` (besucht am 30.06.2007), 2000.

Burrows, John: Word-patterns and storyshapes: The statistical analysis of narrative style. In: *Literary and Linguistic Computing*, Band 2(2):S. 61–70, 1987.

Burrows, John: „An ocean where each kind...": Statistical analysis and some major determinants of literary style. In: *Computers and the Humanities*, Band 23(4):S. 309–321, 1988.

Burrows, John: Not unless you ask nicely: The interpretative nexus between analysis and information. In: *Literary and Linguistic Computing*, Band 7(2):S. 91–109, 1992.

Burrows, John: „Delta": a measure of stylistic difference and a guide to likely authorship. In: *Literary and Linguistic Computing*, Band 17(3):S. 267–287, 2002.

Casella, George und George, Edward I.: Explaining the Gibbs sampler. In: *The American Statistician*, Band 46(3):S. 167–174, 1992.

Ćavar, Damir, Herring, Joshua, Ikuta, Toshikazu, Rodrigues, Paul, und Schrementi, Giancarlo: On induction of morphology grammars and its role in bootstrapping. In: Jäger, Gerhard, Monachesi, Paola, Penn, Gerald und Wintner, Shuly (Hg.) *Proceedings of Formal Grammar 2004*. 2004, S. 47–62.

Chaski, Carole: Empirical evaluations of language-based author identification techniques. In: *Forensic Linguistics*, Band 81:S. 1–65, 2001.

Chaski, Carole: Who's at the keyboard: Authorship attribution in digital evidence investigations. In: *International Journal of Digital Evidence*, Band 4(1), 2005.

Clark, Alexander Simon: *Unsupervised Language Acquisition: Theory and Practice*. Dissertation, University of Sussex, 2001.

Clement, Ross und Sharp, David: Ngram and Bayesian classification of documents for topic and authorship. In: *Literary and Linguistic Computing*, Band 18(4):S. 423–447, 2003.

Cohen, Jacob: The earth is round ($p < .05$). In: *American Psychologist*, Band 49(12):S. 997–1003, 1994.

Cohen, Paul, Adams, Niall und Heeringa, Brent: Voting experts: An unsupervised algorithm for segmenting sequences. In: *Intelligent Data Analysis*, Band 11(6):S. 607–625, 2007.

Corpas, Gloria, Mitkov, Ruslan, Afzal, Naveed und Pekar, Viktor: Translation universals: Do they exist? a corpus-based NLP study of convergence and simplification. In: *Proceedings of the Eighth Conference of the Association for Machine Translation in the Americas (AMTA-08)*. 2008.

Creutz, Mathias: Unsupervised segmentation of words using prior distributions of morph length and frequency. In: *Proc. ACL'03*. Sapporo, Japan, 2003, S. 280–287.

Creutz, Mathias und Lagus, Krista: Unsupervised discovery of morphemes. In: *Proceedings of the Workshop on Morphological and Phonological Learning of the Association for Computational Linguistics (ACL'02)*. 2002, S. 21–30.

Creutz, Mathias und Lagus, Krista: Induction of a simple morphology for highly inflecting languages. In: *Proceedings of the 7th Meeting of the ACL Special Interest Group in Computational Phonology (SIGPHON)*. 2004, S. 43–51.

Creutz, Mathias und Lagus, Krista: Inducing the morphological lexicon of a natural language from unannotated text. In: *Proceedings of the International and Interdisciplinary Conference on Adaptive Knowledge Representation and Reasoning (AKRR'05)*. 2005a, S. 106–113.

Creutz, Mathias und Lagus, Krista: Unsupervised morpheme segmentation and morphology induction from text corpora using morfessor 1.0. Publications in Computer and Information Science Report A81, Helsinki University of Technology, 2005b.

Creutz, Mathias und Lagus, Krista: Unsupervised models for morpheme segmentation and morphology learning. In: *ACM Trans. Speech Lang. Process.*, Band 4(1):S. 1–34, 2007.

Dai, Guangrong und Xiao, Zhonghua: „Source Language Shining Through“ in translational language: A corpus-based study of Chinese translation of English passives. In: *Translation Quarterly*, Band 62:S. 85–107, 2011.

Daniels, Peter T. und Bright, William: *The World's Writing Systems*. Oxford University Press, Oxford, 1996.

de Marcken, Carl G.: *Unsupervised Language Acquisition*. Dissertation, Massachusetts Institute of Technology, 1996.

de Morgan, Sophia Elisabeth: *Memoir of Augustus de Morgan by his Wife Sophia Elisabeth de Morgan With Selections From His Letters*. Longmans, Green, and Co., London, 1882.

Dejean, Hervé: Morphemes as necessary concept for structures discovery from untagged corpora. In: *Proceedings of the Workshop on Paradigms and Grounding in Natural Language Learning (CoNLL'98)*. 1998, S. 295–299.

Diederich, Joachim, Kindermann, Jörg, Leopold, Edda und Paass, Gerhard: Authorship attribution with support vector machines. In: *Applied Intelligence*, Band 19(1–2):S. 109–123, 2003.

Ebeling, Werner und Neiman, Alexander: Long-range correlations between letters and sentences in texts. In: *Physica A*, Band 215:S. 233–242, 1995.

Ebeling, Werner und Pöschel, Thorsten: Entropy and long range correlations in literary English. In: *Europhysics Letters*, Band 26(2):S. 241–246, 1994.

Ebeling, Werner, Pöschel, Thorsten und Albrecht, Karl-Friedrich: Entropy, transinformation and word distribution of information-carrying sequences. In: *Int. J. Bifurcation & Chaos*, Band 5:S. 51–61, 1995.

Estival, Dominique, Gaustad, Tanja, Hutchinson, Ben, Pham, Son Bao und Radford, Will: Author profiling for English and Arabic emails. URL `http://hdl.handle.net/2123/5839` (besucht am 13.10.2012), 2008.

Feng, Haodi, Chen, Kang, Kit, Chunyu und Deng, Xiaotie: Unsupervised segmentation of Chinese corpus using accessor variety. In: *Proceedings of IJCNLP 2004*. Springer, Hainan Island, China, 2004, S. 694–703.

Fontane, Theodor: Effi Briest. URL `http://www.gutenberg.org/ebooks/5323` (besucht am 15.10.2012), 2004.

Forsyth, Richard S.: Towards a text benchmark suite. In: *Joint International Conference of the Association for Computers and the Humanities and the Association for Literary and Linguistic Computing (ACH/ALLC 1997)*. Kingston, ON, 1997.

Forsyth, Richard S., Holmes, David I. und Tse, Emily K.: Cicero, Sigonio, and Burrows: investigating the authenticity of the Consolatio. In: *Literary and Linguistic Computing*, Band 14:S. 375–400, 1999.

Francis, W.Nelson und Kucera, Henry: Brown Corpus Manual. 1967. Besucht am 19.06.2012, URL `http://khnt.aksis.uib.no/icame/manuals/brown/`.

Fung, Glenn: The disputed federalist papers: SVM feature selection via concave minimization. In: *TAPIA '03: Proceedings of the 2003 conference on Diversity in computing*. ACM Press, New York, NY, USA, 2003, S. 42–46.

Gamon, Michael: Linguistic correlates of style: authorship classification with deep linguistic analysis features. In: *Proc. 20th Int. Conf. Computational Linguistics (COLING)*. Geneva, 2004, S. 611–617.

Gaussier, Eric: Unsupervised learning of derivational morphology from inflectional lexicons. In: *Proceedings of the Workshop On Unsupervised Learning In Natural Language Processing*. 1999, S. 24–30.

Gellerstam, Martin: Translationese in swedish novels translated from English. In: Wollin, Lars und Lindquist, Hans (Hg.) *Translation Studies in Scandinavia*, CWK Gleerup, Lund, S. 88–95. 1986.

Gilden, David L., Thornton, Thomas L. und Mallon, Marc W.: $1/f$ noise in human cognition. In: *Science*, Band 267:S. 1837–1839, 1995.

Golcher, Felix: Statistische Aspekte von Suffixbäumen natürlichsprachiger Texte. URL `http://www.hu-berlin.de/~golcherf/suffix.htm` (besucht am 13.10.2012), 2005. Abschlussarbeit für den Aufbaustudiengang Computerlinguistik des Centrum für Informations- und Sprachverarbeitung der Universität München.

Golcher, Felix: Statistical text segmentation with partial structure analysis. In: *Proceedings of KONVENS 2006*. Universität Konstanz, Konstanz, 2006, S. 44–51.

Golcher, Felix: A new text statistical measure and its application to stylometry. In: Davies, Matthew, Rayson, Paul, Hunston, Susan und Danielsson, Pernilla (Hg.) *Corpus Linguistics 2007*. University of Birmingham, Birmingham, 2007a.

Golcher, Felix: A stable statistical constant specific for human language texts. In: *Recent Advances in Natural Language Processing 2007 (RANLP-07)*. Bulgarian Academy of Sciences, Sofia, 2007b.

Golcher, Felix und Reznicek, Marc: Stylometry and the interplay of topic and l1 in the different annotationlayers in the falko corpus. In: Zeldes, Amir und Lüdeling, Anke

(Hg.) *Proceedings of Quantitative Investigations in Theoretical Linguistics 4 (QITL-4)*. Humboldt-Universität zu Berlin, Berlin, 2011.

Goldsmith, John: Unsupervised learning of the morphology of a natural language. In: *Comput. Linguist.*, Band 27(2):S. 153–198, 2001.

Goldsmith, John: Segmentation and morphology. In: Clark, Alex, Fox, Chris und Lappin, Shalom (Hg.) *The Handbook of Computational Linguistics and Natural Language Processing*, Blackwell, S. 364–393. 2010.

Goldsmith, John, Higgins, Derrick und Soglasnova, Svetlana: Automatic language-specific stemming in information retrieval. In: Peters, Carol (Hg.) *Cross-Language Information Retrieval and Evaluation*, Springer, Berlin, Heidelberg, Band 2069 von *Lecture Notes in Computer Science*, S. 273–283. 2001.

Goldwater, Sharon, Griffiths, Thomas L. und Johnson, Mark: Interpolating between types and tokens by estimating power-law generators. In: *In Advances in Neural Information Processing Systems 18*. 2006, S. 18.

Goldwater, Sharon, Griffiths, Thomas L. und Johnson, Mark: A Bayesian framework for word segmentation: Exploring the effects of context. In: *Cognition*, Band 112(1):S. 21–54, 2009.

Goldwater, Sharon J.: *Nonparametric Bayesian Models of Lexical Acquisition*. Dissertation, Brown University, 2007.

González, Antoni Oliver: *Adquisició d'informació lèxica i morfosintàactica a partir de corpus sense anotar: aplicació al rus i al croat*. Dissertation, Universitat de Barcelona, 2004.

Goodman, Joshua: Extended comment on language trees and zipping. URL `http://front.math.ucdavis.edu/0202.0383` (besucht am 14.10.2012), 2002.

Granados, Ana, Cebrián, Manuel, Camacho, David und Rodríguez, Francisco B.: Evaluating the impact of information distortion on normalized compression distance. In: Barbero, Angela I. (Hg.) *ICMCTA*. Springer, Berlin, 2008, Band 5228 von *Lecture Notes in Computer Science*, S. 69–79.

Granger, Sylviane: The international corpus of learner English: A new resource for foreign language learning and teaching and second language acquisition research. In: *Tesol Quarterly*, Band 37(3):S. 538–546, 2003.

Grewendorf, Günther, Hamm, Fritz und Sternefeld, Wolfgang: *Sprachliches Wissen: Eine Einführung in moderne Theorien der grammatischen Beschreibung*. Suhrkamp, Frankfurt am Main, 1987.

Gries, Stefan Th.: Null-hypothesis significance testing of word frequencies: a follow-up on Kilgarriff. In: *Corpus Linguistics and Linguistic Theory*, Band 1–2:S. 277–294, 2005.

Gries, Stefan Th.: Exploring variability within and between corpora: some methodological considerations. In: *Corpora*, Band 1(2):S. 109–151, 2006.

Gries, Stefan Th.: Dispersions and adjusted frequencies in corpora. In: *International Journal of Corpus Linguistics*, Band 13(4):S. 403–437, 2008.

Grieve, Jack: Quantitative authorship attribution: An evaluation of techniques. In: *Literary and Linguistic Computing*, Band 22(3):S. 251–270, 2007.

Grimm, Jacob und Grimm, Wilhelm: *Deutsches Wörterbuch*, Band 24. Deutscher Taschenbuchverlag, 1984. Fotomech. Nachdr. d. Erstausg. 1936.

Gusfield, Dan: *Algorithms on strings, trees, and sequences: computer science and computational biology*. Cambridge University Press, 1997.

Hafer, Margaret A. und Weiss, Stephen F.: word segmentation by letter successor varieties. In: *Inform. Stor. Retr.*, Band 10:S. 371–385, 1974.

Hamilton, Alexander, Jay, John und Madison, James: The federalist papers. URL `http://www.gutenberg.org/etext/18` (besucht am 14.10.2012), 2004.

Hammarström, Harald: A naive theory of morphology and an algorithm for extraction. In: *SIGPHON-06*. 2006.

Hammarström, Harald: *Unsupervised Learning of Morphology and the Languages of the World*. Dissertation, Chalmers University, 2009.

Harris, Zellig S.: From phoneme to morpheme. In: *Language*, Band 31(2):S. 190–222, 1955. Reprinted in Hiż (1970).

Harris, Zellig S.: Morpheme boundaries within words: Report on a computer test. In: *Transformations and Discourse Analysis Papers*, Band 73, 1967. Reprinted in Hiż (1970).

Harris, Zellig S.: Recurrent dependence process: Morphemes by phoneme neighbours. In: *Mathematical structures of language*, Interscience, New York, Band 21 von *Interscience tracts in pure and applied mathematics*, S. 24–28. 1968.

Hilberg, Wolfgang: The well-known lower bound of information in written language - is it a misinterpretation of shannon's experiments? In: *Frequenz*, Band 44:S. 243–248, 1990.

Hirst, Graeme und Feiguina, Ol'ga: Bigrams of syntactic labels for authorship discrimination of short texts. In: *Literary and Linguistic Computing*, Band 22(4):S. 405–417, 2007.

Hiż, Henry (Hg.): *Papers in Structural and Transformational Linguistics*. Dordrecht, Holland, 1970.

Hockett, Charles F.: Problems of morphemic analysis. In: *Language*, Band 23:S. 321–43, 1947.

Holmes, David I.: Authorship attribution. In: *Computers and the Humanities*, Band 28:S. 87–106, 1994.

Holmes, David I.: The evolution of sylometry in humanities scholarship. In: *Literary and Linguistic Computing*, Band 13(3):S. 111–127, 1998.

Holmes, David I. und Forsyth, Richard S.: The federalist revisited: New directions in authorship attribution. In: *Literary and Linguistic Computing*, Band 10(2):S. 111–127, 1995.

Holmes, David. I., Robertson, Michael und Paez, Roxanna: Stephen Crane and the New-York Tribune: A case study in traditional and non-traditional authorship attribution. In: *Computers and the Humanities*, Band 35(3):S. 315–331, 2001.

Hoover, David L.: Another perspective on vocabulary richness. In: *Computers and the Humanities*, Band 37(2):S. 151–78, 2003a.

Hoover, David L.: Multivariate analysis and the study of style variation. In: *Literary and Linguistic Computing*, Band 18(4):S. 341–360, 2003b.

Hothorn, Torsten und Hornik, Kurt: exactranktests: Exact distributions for rank and permutation tests. URL `http://CRAN.R-project.org/package=exactRankTests` (besucht am 14.10.2012), 2011. R package version 0.8-22.

Ilisei, Iustina, Inkpen, Diana, Pastor, Gloria Corpas und Mitkov, Ruslan: Identification of translationese: A machine learning approach. In: *Computational Linguistics and Intelligent Text Processing*, Springer, Berlin, Heidelberg, Band 6008 von *Lecture Notes in Computer Science*, S. 503–511. 2010.

Jockers, Matthew L. und Witten, Daniela M.: A comparative study of machine learning methods for authorship attribution. In: *Literary and Linguistic Computing*, Band 25(2):S. 215–223, 2010.

Johnson, Mark: Using adaptor grammars to identify synergies in the unsupervised acquisition of linguistic structure. In: *Proceedings of ACL-08: HLT*. Association for Computational Linguistics, Columbus, Ohio, 2008, S. 398–406.

Johnson, Mark, Griffiths, Thomas L. und Goldwater, Sharon: Adaptor Grammars: A Framework for Specifying Compositional Nonparametric Bayesian Models. In: *Advances in Neural Information Processing Systems 19*. MIT Press, 2007, S. 641–648.

JojoWong, Sze-Meng und Dras, Mark: Contrastive analysis and native language identification. In: Pizzato, Luiz Augusto und Schwitter, Rolf (Hg.) *Australasian Language Technology Association Workshop 2009*. University of New South Wales, Sydney, Australia, 2009, S. 53–61.

Juola, Patrick: Ad-hoc authorship attribution competition. In: *Proceedings 2004 Joint International Conference of the Association for Literary and Linguistic Computing and the Association for Computers and the Humanities (ALLC/ACH 2004)*. Göteborg, Sweden, 2004, S. 175–176.

Juola, Patrick: Authorship attribution. In: *Foundations and Trends in Information Retrieval*, Band 1(3):S. 233–334, 2006a.

Juola, Patrick: Questioned electronic documents: Empirical studies in authorship attribution. In: Olivier und Shenoi (Hg.) *Research Advances in Digital Forensics II*, Springer, Heidelberg. 2006b.

Juola, Patrick und Baayen, R. Harald: A controlled-corpus experiment in authorship identification by cross-entropy. In: *Literary and Linguistic Computing*, Band 20:S. 59–67, 2005.

Juola, Patrick, Sofko, John und Brennan, Patrick: A prototype for authorship attribution studies. In: *Literary and Linguistic Computing*, Band 21(2):S. 169–178, 2006.

Kaneko, Kunihiko und Li, Wentian: Long-range correlation and partial $1/f$ α spectrum in a noncoding DNA sequence. In: *Europhysics Letters*, Band 17(7):S. 655–660, 1992.

Kant, Immanuel: Kritik der reinen Vernunft. URL `http://www.gutenberg.org/dirs/etext04/8ikc110.txt` (besucht am 15.10.2012), 2004.

Kello, Christopher T., Brown, Gordon D. A., i Cancho, Ramon Ferrer, Holden, John G., Linkenkaer-Hansen, Klaus, Rhodes, Theo und Orden, Guy C. Van: Scaling laws in cognitive sciences. In: *Trends in Cognitive Science*, Band 14(5):S. 223–232, 2010.

Keselj, Vlado und Cercone, Nick: CNG method with weighted voting. In: *Ad-hoc Authorship Attribution Contest. ACH/ALLC 2004*. 2004. Konferenzposter.

Khmelev, Dmitry V. und Teahan, William J.: Comment on "language trees and zipping". In: *Phys. Rev. Lett.*, Band 90(8):S. 089803, 2003.

Kilgarriff, Adam: Language is never, ever, ever, random. In: *Corpus Linguistics and Linguistic Theory*, Band 1(2):S. 263–276, 2005.

Klemperer, Victor: *LTI. Notizbuch eines Philologen.* Reclam Leipzig, 1975.

Koppel, Moshe, Schler, Jonathan und Argamon, Shlomo: Computational methods in authorship attribution. In: *JASIST*, Band 60(1):S. 9–26, 2009.

Koppel, Moshe, Schler, Jonathan und Bonchek-Dokow, Elisheva: Measuring differentiability: Unmasking pseudonymous authors. In: *Journal of Machine Learning Research*, Band 8:S. 1261–1276, 2007.

Koppel, Moshe, Schler, Jonathan und Zigdon, Kfir: Determining an author's native language by mining a text for errors. In: *Proceedings of KDD '05*. Chicago IL, 2005, S. 624–628.

Koppel, Moshe, Schler, Jonathan und Zigdon, Kfir: Automatically determining an anonymous author's native language. In: Mehrotra, Sharad, Zeng, Daniel D. und Chen, Hsinchun (Hg.) *Intelligence and Security Informatics*, Springer, Berlin, Heidelberg, Lecture Notes in Computer Science, S. 209–217. 2006.

Kriz, Thomas A. und Talacko, Joseph V.: Equivalence of the maximum likelihood estimator to a minimum entropy estimator. In: *Trabajos de Estadística y de Investigación Operativa*, Band 1–2:S. 55–65, 1968.

Kuĉera, Henry und Francis, W. Nelson: *Computational Analysis of Present-Day American English*. Brown University Press, Providence, RI, USA, 1967.

Kurimo, Mikko, Virpioja, Sami und Turunen, Ville T. (Hg.): *Proceedings of the Morpho Challenge 2010 Workshop*, TKK Reports in Information and Computer Science. Helsinki University of Technology, Department of Information and Computer Science, 2010.

Kurokawa, David, Goutte, Cyril und Isabelle, Pierre: Automatic detection of translated text and its impact on machine translation. In: *MT Summit XII: proceedings of the twelfth Machine Translation Summit*. 2009, S. 81–88.

Laviosa, Sara: Core patterns of lexical use in a comparable corpus of english narrative prose. In: *The Corpus-Based Approach*, Les Presses de L'Université de Montréal, Montréal, S. 557–570. 1998.

Laviosa, Sara: *Corpus-based Translation Studies. Theory, Findings, Applications.* Rodopi, Amsterdam, 2002.

Levitin, Lev B. und Reingold, Zeev: Entropy of natural languages: Theory and experiment. In: *Chaos, Solitons & Fractals*, Band 4(5):S. 709 – 743, 1994.

Lieber, Rochelle und Mugdan, Joachim: Internal structure of words. In: Booij, Geert, Lehmann, Christian und Mugdan, Joachim (Hg.) *Morphology*, Walter de Gruyter, HSK, S. 404–416. 2000.

Lüdeling, Anke: Coding word-formation morphology in computational dictionaries. In: Gouws, Rufus H., Heid, Ulrich, Schweickard, Wolfgang und Wiegand, Herbert Ernst (Hg.) *Dictionaries. An International Encyclopedia of Lexicography*, Mouton de Gruyter, Berlin. erscheint.

Lüdeling, Anke, Doolittle, Seanna, Hirschmann, Hagen, Schmidt, Karin und Walter, Maik: Das Lernerkorpus Falko. In: *Deutsch als Fremdsprache*, Band 2:S. 67–73., 2008.

MacWhinney, Brian und Snow, Catherine: The child language data exchange system. In: *Journal of Child Language*, Band 12:S. 271–296, 1985.

Manning, Christopher D. und Schütze, Hinrich: *Foundations of statistical natural language processing*. MIT Press, Cambridge, MA, 1999.

Martindale, Colin und McKenzie, Dean: On the utility of content analysis in author attribution: The 'federalist'. In: *Computers and the Humanities*, Band 29:S. 259–270, 1995.

Mendenhall, Thomas Corwin: The characteristic curves of composition. In: *Science*, Band IX:S. 237–249, 1887.

Mendenhall, Thomas Corwin: A mechanical solution to a literary problem. In: *Popular Science Monthly*, Band 9:S. 97–110, 1901.

Merriam, Thomas V. N.: Marlowe's hand in *Eduard III.* In: *Literary and Linguistic computing*, Band 8(2):S. 59–72, 1993.

Merriam, Thomas V. N.: Edward III. In: *Literary and Linguistic Computing*, Band 15(2):S. 157–186, 2000.

Merriam, Thomas V. N. und Matthews, Robert A. J.: Neural computation in stylometry II: An application to the works of shakespeare and marlowe. In: *Literary and Linguistic Computing*, Band 9(1):S. 1–6, 1994.

Milne, Alan A.: *Winnie-the-Pooh.* Methuen & Co. Ltd., London, 1926.

Mochihashi, Daichi und Sumita, Eiichiro: The infinite markov model. In: Platt, John C., Koller, Daphne, Singer, Yoram und Roweis, Sam T. (Hg.) *Advances in Neural Information Processing Systems 20 (NIPS 2007).* MIT Press, Cambridge, MA, 2008, S. 1017–1024.

Mochihashi, Daichi, Yamada, Takeshi und Ueda, Naonori: Bayesian unsupervised word segmentation with nested Pitman-Yor language modeling. In: *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP.* Association for Computational Linguistics, Suntec, Singapore, 2009, S. 100–108.

Mollet, Eugène, Wray, Alison, Fitzpatrick, Tess, Wray, Naomi R. und Wright, Margaret J.: Choosing the best tools for comparative analyses of texts. In: *Journal of Corpus Linguistics*, Band 15(4):S. 429–473, 2010.

Montemurro, Marcelo A. und Pury, Pedro A.: Long-range fractal correlations in literary corpora. In: *Fractals*, Band 10(4):S. 451–461, 2002.

Mosteller, Frederick und Wallace, David L.: *Inference and disputed authorship, the Federalist.* Adison Wesley, Reading, MA, 1964.

Mosteller, Frederick und Wallace, David L.: *Applied Bayesian and Classical Inference: The Case of the Federalist Papers.* Springer series in statistics. Springer, New York, Heidelberg, 1984.

Mugdan, Joachim: Morphological Units. In: Asher, Ronald E. (Hg.) *The Encyclopedia of Language and Linguistics*, Pergamon Press, Tokyo, S. 2543–2553. 1994.

Neuvel, Sylvain und Fulop, Sean A.: Unsupervised learning of morphology without morphemes. In: *Proceedings of the ACL-02 workshop on Morphological and phonological learning - Volume 6*. Association for Computational Linguistics, Stroudsburg, PA, 2002, S. 31–40.

Oberlander, Jon und Brew, Chris: Stochastic text generation. In: *Philosophical Transactions of the Royal Society of London*, Band 358(1769):S. 1373–1385, 2000.

Peng, Chung Kang, Buldyrev, Sergej V., Goldberger, Ary L., Havlin, Shlomo, Sciortino, Francesco, Simons, Michael und Stanley, H. Eugene: Long-range correlations in nucleotide sequences. In: *Nature*, Band 356(6365):S. 168–170, 1992.

Peng, Fuchun und Schuurmans, Dale: Self-supervised Chinese word segmentation. In: Hoffmann, Frank, Hand, DavidJ., Adams, Niall, Fisher, Douglas und Guimaraes, Gabriela (Hg.) *Advances in Intelligent Data Analysis*. 2001, Band 2189 von *Lecture Notes in Computer Science*, S. 238–247.

Pinheiro, José C. und Bates, Douglas M.: *Mixed-Effects Models in S and S-PLUS*. Springer, Berlin, Heidelberg, 2000.

Pinheiro, José C., Bates, Douglas M., DebRoy, Saikat, Sarkar, Deepayan und R Development Core Team: *nlme: Linear and Nonlinear Mixed Effects Models*, 2011. R package version 3.1-102.

Press, William H.: Flicker noise in astronomy and elsewhere. In: *Comments Astrophys*, Band 7:S. 103–119, 1978.

Rissanen, Jorma: *Stochastic Complexity in Statistical Inquiry*. World Scientific Publishing Co, Singapore, 1989.

Roark, Brian und Sproat, Richard: Machine learning of morphology. In: *Computational approaches to morphology and syntax*, Oxford University Press, Oxford, Band 4 von *Oxford surveys in syntax and morphology*, S. 116–136. 2007.

Rowohlt, Harry: Warum ein Bär den Honig mag. URL `http://www.nordbayern.de/nuernberger-zeitung/cs34-7-warum-ein-bar-den-honig-mag-1.710345` (besucht am 3.11.2012), 2005. Harry Rowohlt im Gespräch mit der Nürberger Zeitung.

Rudman, Joseph: The state of authorship attribution studies: Some problems and solutions. In: *Computers and the Humanities*, Band 31:S. 351–365, 1998.

Rybski, Diego, Buldyrev, Sergey V., Havlin, Shlomo, Liljeros, Fredrik und Makse, Hernán A.: Scaling laws of human interaction activity. In: *Proceedings of the National Academy of Sciences*, Band 106(31):S. 12640–12645, 2009.

Saffran, Jenny R., Aslin, Richard N. und Newport, Elissa L.: Statistical learning by 8-month old infants. In: *Science*, Band 274:S. 1926–1928, 1996a.

Saffran, Jenny R., Newport, Elissa L. und Aslin, Richard N.: Word segmentation: The role of distributional cues. In: *Journal of Memory and Language*, Band 35(4):S. 606–621, 1996b.

Salton, Gerard und McGill, Michael J.: *Introduction to modern information retrieval*. McGraw-Hill, New York, 1983.

Salton, Gerard, Wong, Andrew K. C. und Yang, Chung-Shu: A vector space model for automatic indexing. In: *Communications of the ACM*, Band 18(11):S. 613–620, 1975.

Say, Bilge, Zeyrek, Deniz, Oflazer, Kemal und Özge, Umut: Development of a corpus and a treebank for present-day written Turkish. In: İmer, Kamile und Doğan, Gürkan (Hg.) *Proceedings of the Eleventh International Conference of Turkish Linguistics*. Eastern Mediterranean University, Famagusta, Zypern, 2002, S. 183–192.

Schenkel, Alain, Zhang, Jun und Zhang, Yi-Cheng: Long range correlations in human writings. In: *Fractals*, Band 1(1):S. 47–57, 1993.

Schmid, Helmut: Probabilistic part-of-speech tagging using decision trees. In: *Proceedings of the International conference on New Methods in Language Processing*. University of Manchester, Manchester, 1994, S. 44–49.

Schölkopf, Bernhard und Smola, Alexander J.: *Learning with Kernels: Support Vector Machines, Regularization, Optimization and Beyond*. MIT Press, Cambridge, MA, 2002.

Schone, Patrick und Jurafsky, Daniel: Knowledge-free induction of morphology using latent semantic analysis. In: *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning - Volume 7*. Association for Computational Linguistics, Stroudsburg, PA, USA, 2000, S. 67–72.

Schone, Patrick und Jurafsky, Daniel: Knowledge-free induction of inflectional morphologies. In: *Proceedings of the North American Chapter of the ACL*. Association for Computational Linguistics, Stroudsburg, PA, 2001, S. 183–191.

Shannon, Claude E.: A mathematical theory of communication. In: *Bell System technical journal*, Band 27:S. 379–423, 1948.

Shannon, Claude E.: Prediction and entropy of printed english. In: *Bell System Technical Journal*, Band 30:S. 50–64, 1951.

Sharma, Utpal, Kalita, Jugal und Das, Rajib: Unsupervised learning of morphology for building lexicon for a highly inflectional language. In: *Proceedings of the ACL-02 workshop on Morphological and phonological learning - Volume 6*. Association for Computational Linguistics, Stroudsburg, PA, USA, 2002, S. 1–10.

Sherman, Lucius Adelno: principle of sentence length as an indicator of style and attribution. 1888. Die genaue Referenz war nicht ermittelbar.

Shlesinger, Miriam: Towards a definition of interpretese: An intermodal, corpus-based study. In: Hansen, Gyde, Chesterman, Andrew und Gerzymisch-Arbogast, Heidrun (Hg.) *Efforts and Models in Interpreting and Translation Research: A tribute to Daniel Gile*, John Benjamins, Amsterdam, S. 237–253. 2009.

Singh, Anil Kumar und Gorla, Jagadeesh: Identification of languages and encodings in a multilingual document. In: *Proceedings of the 3rd ACL SIGWAC Workshop on Web As Corpus*. Louvain-la-Neuve, Belgium, 2007.

Snover, Matthew G. und Brent, Michael R.: A bayesian model for morpheme and paradigm identification. In: *Proceedings of the ACL'01*. Morgan Kaufmann Publishers, San Francisco, CA, 2001, S. 482–490.

Snover, Matthew G., Jarosz, Gaja E. und Brent, Michael R.: Unsupervised learning of morphology using a novel directed search algorithm: taking the first step. In: *Proceedings of the ACL-02 workshop on Morphological and phonological learning - Volume 6*. Association for Computational Linguistics, Stroudsburg, PA, USA, 2002, S. 11–20.

Stamatatos, Efstathios: A survey of modern authorship attribution methods. In: *J. Am. Soc. Inf. Sci. Technol.*, Band 60(3):S. 538–556, 2009.

Teahan, Wiliam J.: Text classification and segmentation using minimum cross-entropy. In: *Proceedings of RIAO–00, 6th International conference „Recherche d'Information Assisté par Ordinateur"*. College de France, Paris, 2000, Band 2, S. 943–961.

Teh, Yee Whye: A hierarchical bayesian language model based on Pitman-Yor processes. In: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, 2006, S. 985–992.

Teich, Elke: *Cross-linguistic Variation in System and Text*. Mouton de Gruyter, Berlin, 2003.

Tepper, Michael und Xia, Fei: Inducing morphemes using light knowledge. In: *ACM Transactions on Asian Language Information Processing (TALIP)*, Band 9(1):S. 1–38, 2010.

Toury, Gideon: *Descriptive Translation Studies and Beyond*. John Benjamins, Amsterdam, 1995.

Trost, Harald: Morphology. In: *The Oxford Handbook of Computational Linuistics*, Oxford University Press, Oxford, S. 25–47. 2003.

Tsur, Oren und Rappoport, Ari: Using classifier features for studying the effect of native language choice of written second language words. In: *Proceedings of the Workshop on Cognitive Aspects of Computational Language Acquisition*. Association for Computational Linguistics, Stroudsburg, PA, 2007, S. 9–16.

Tweedie, Fiona und Baayen, R. Harald: How variable may a constant be? measures of lexical richness in perspective. In: *Computers and the Humanities*, Band 32:S. 323–352, 1998.

Tweedie, Fiona J., Singh, Sameer und Holmes, David I.: Neural network applications in stylometry: the *Federalist papers*. In: *Computers and the Humanities*, Band 30:S. 1–10, 1996.

Ukkonen, Esko: On-line construction of suffix-trees. In: *Algorithmica*, Band 14(3):S. 249–260, 1995.

Usatenko, Oleg V. und Yampol'skii, Valery Aleksandrovich: Binary n-step markov chains and long-range correlated systems. In: *Phys. Rev. Lett.*, Band 90(11):S. 110601, 2003.

Uzuner, Özlem und Katz, Boris: A comparative study of language models for book and author recognition. In: *Proceedings of the Second international joint conference on Natural Language Processing*. Springer, Berlin, Heidelberg, 2005, IJCNLP'05, S. 969–980.

van Halteren, Hans: Source language markers in EUROPARL translations. In: *COLING '08: Proceedings of the 22nd International Conference on Computational Linguistics*. Association for Computational Linguistics, Morristown, NJ, USA, 2008, S. 937–944.

Venkataraman, Anand: A statistical model for word discovery in transcribed speech. In: *Computational Linguistics*, Band 27(3):S. 351–372, 2001.

Vicente, Kim J. und Torenvliet, Gerard L.: The earth is spherical ($p < 0.05$): alternative methods of statistical inference. In: *Theoritical Issues in Ergonomics Science*, Band 1:S. 248 – 271, 2000.

Voss, Richard F.: Evolution of long-range fractal correlations and $1/f$ noise in DNA base sequences. In: *Phys. Rev. Lett.*, Band 68:S. 3805–3808, 1992.

Voss, Richard F. und Clarke, John: „$1/f$ noise“ in music and speech. In: *Nature*, Band 258:S. 317–318, 1975.

Weiner, Peter: Linear pattern matching algorithms. In: *Proceedings of the 14th Annual Symposium on Switching and Automata Theory (swat 1973)*. IEEE Computer Society, Washington, DC, 1973, S. 1–11.

Wicentowski, Richard: *Modeling and learning multilingual inflectional morphology in a minimally supervised framework*. Dissertation, Johns Hopkins University, 2002.

Wikipedia-Mitarbeiter: Federalist papers. URL `http://en.wikipedia.org/wiki/Federalist_papers` (besucht am 14.10.2012), 2001.

Wikipedia-Mitarbeiter: Großschreibung. URL `http://de.wikipedia.org/wiki/Gro%C3%9Fschreibung` (besucht am 15.10.2012), 2005.

Wittgenstein, Ludwig: *Philosophische Untersuchungen.* Wissenschaftliche Buchgesellschaft, Frankfurt, 2001.

Wulff, Stefanie: *Rethinking Idiomaticity.* Continuum, London, New York, 2009.

Wurzel, Wolfgang Ullrich: *Flexionsmorphologie und Natürlichkeit.* Nummer 21 in Studia grammatica. Akademie-Verlag, Berlin, 1984.

Xu, Jia, Gao, Jianfeng, Toutanova, Kristina und Ney, Hermann: Bayesian semi-supervised chinese word segmentation for statistical machine translation. In: *Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1.* Association for Computational Linguistics, Stroudsburg, PA, USA, 2008, S. 1017–1024.

Yamasaki, Kazuko, Muchnik, Lev, Havlin, Shlomo, Bunde, Armin und Stanley, H. Eugene: Scaling and memory in volatility return intervals in financial markets. In: *Proceedings of the National Academy of Sciences of the United States of America*, Band 102(26):S. 9424–9428, 2005.

Yarowsky, David und Wicentowski, Richard: Minimally supervised morphological analysis by multimodal alignment. In: *Proceedings of ACL-2000.* Association for Computational Linguistics, Stroudsburg, PA, 2000, S. 207–216.

Yu, Bei: An evaluation of text classification methods for literary study. In: *Literary and Linguistic Computing*, Band 23(3):S. 327–343, 2008.

Yu, Hua: Unsupervised word induction using MDL criterion. In: *ISCSL.* Beijing, 2000.

Yule, George Udny: On sentence length as a statistical characteristic of style in prose with application to two cases of disputed authorship. In: *Biometrika*, Band 30:S. 363–390, 1938.

Yule, George Udny: *The statistical study of literary vocabulary.* Cambridge University Press, Cambridge, 1944.

Yule, George Udny: *The statistical study of literary vocabulary.* Archon Books, Hamden, CT, 1968. Reprint of Yule (1944).

Zhang, Dell und Lee, Wee Sun: Extracting key-substring-group features for text classification. In: *Proceedings of the 12th Annual SIGKDD International Conference on Knowledge Discovery and Data Mining.* ACM, New York, NY, 2006, S. 474–483.

Zhao, Ying und Zobel, Justin: Effective and scalable authorship attribution using function words. In: Lee, Gary Geunbae, Yamada, Akio, Meng, Helen und Myaeng, Sung-Hyon (Hg.) *2nd Asian Information Retrieval Symposium.* Springer, Berlin, Heidelberg, 2005, Band 3689 von *Lecture Notes in Computer Science*, S. 174–190.

Zigdon, Kfir: *Automatically determining an author's native language.* Diplomarbeit, Dept. of Computer Science, Bar-Ilan University, 2005.

Zipf, George Kingsley: *Human Behavior and The Principle of Least Effort.* Hafner Publishing Company, New York, London, 1949.

Zuur, Alain F., Ieno, Elena N., Walker, Neil J., Saveliev, Anatoly A. und Smith, Graham M.: *Mixed Effects Models and Extensions in Ecology with R.* Springer, Berlin, Heidelberg, 2009.

Abbildungsverzeichnis

Tabellenverzeichnis

Selbstständigkeitserklärung zur Dissertation

Humboldt-Universität zu Berlin
Philosophische Fakultät II

Felix Golcher (Matrikelnummer 506583)

Ich erkläre ausdrücklich, dass es sich bei der von mir eingereichten schriftlichen Arbeit mit dem Titel

Wiederholungen in Texten: Segmentieren und Klassifizieren mit vollständigen Substringfrequenzen

um eine von mir selbstständig und ohne fremde Hilfe verfasste Arbeit handelt.

Ich erkläre ausdrücklich, dass ich sämtliche in der oben genannten Arbeit verwendeten fremden Quellen, auch aus dem Internet (einschließlich Tabellen, Grafiken u. Ä.) als solche kenntlich gemacht habe. Insbesondere bestätige ich, dass ich ausnahmslos sowohl bei wörtlich übernommenen Aussagen bzw. unverändert übernommenen Tabellen, Grafiken u. Ä. (Zitaten) als auch bei in eigenen Worten wiedergegebenen Aussagen bzw. von mir abgewandelten Tabellen, Grafiken u. Ä. anderer Autorinnen und Autoren (Paraphrasen) die Quelle angegeben habe.

Mir ist bewusst, dass Verstöße gegen die Grundsätze der Selbstständigkeit als Täuschung betrachtet und entsprechend der Prüfungsordnung und/oder der Allgemeinen Satzung für Studien- und Prüfungsangelegenheiten der HU (ASSP) geahndet werden.

25. Oktober 2012
Felix Golcher